

MEASUREMENT and EVALUATION
in
EDUCATION



THE MACMILLAN COMPANY
NEW YORK • CHICAGO
DALLAS • ATLANTA • SAN FRANCISCO
LONDON • MANILA
BRETT-MACMILLAN LTD.
TORONTO

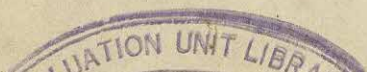
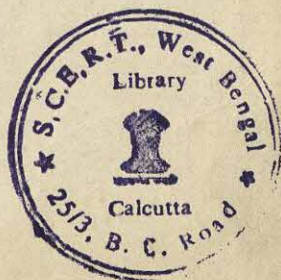
MEASUREMENT and EVALUATION **in** **EDUCATION**

An introduction to its theory and practice
at both the elementary and secondary school levels

JAMES M. BRADFIELD
Professor of Education
Sacramento State College

H. STEWART MOREDOCK
Associate Professor of Mathematics
Sacramento State College

The Macmillan Company • New York



© THE MACMILLAN COMPANY 1957

All rights reserved—no part of this book may be reproduced in any form without permission in writing from the publisher, except by a reviewer who wishes to quote brief passages in connection with a review written for inclusion in magazine or newspaper.

PRINTED IN THE UNITED STATES OF AMERICA

First Printing

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 57-6354

S.C.E.R.T., West Bengal

Date...15-3-85...

Acc. No...3164...

xy

EVALUATION UNIT LIBRARY	
DAVID HARE TRAINING	
COLLEGE	
CALCUTTA-19	
Acc. No.	0037
Date	30 MAR 1965
Call	37126
No.	B799

3164

3164

PREFACE

The concern of this book is what teachers, principals, supervisors, and other educational specialists need to know and do if they are to deal efficiently with measurement and evaluation. Among such items of essential knowledge and performance are thought to be the following:

1. The basic concepts of measurement and evaluation that underlie valid practice.

2. The technical terminology involved.

3. Phenomena that may deserve measurement and their measurable dimensions.

4. The nature of measurement symbols; the many procedures of measurement useful in the schools; and certain statistical ideas and operations important for proper interpretation and use of test results.

5. Standards appropriate to evaluating pupil achievement and efficient ways of reporting evaluations to pupils and parents.

6. How all these things apply to your specialization as to subject, grade, function, etc.

Our treatment of these matters is based on a given rationale of measurement and evaluation and is intentionally developmental and analytic in character. Passages and chapters are interrelated and interdependent. Definitions and principles developed in one chapter are applied in subsequent ones. The first section deals with basic concepts, terminology, and the general features of dimensions, symbols, procedures, statistics, standards, marking, and reporting. In the second section these concepts are applied to school subjects, to intelligence, and to character and personality variables.

Consequently, it is recommended that each of the chapters in Section I be studied carefully and in order, in advance of any reading in Section II. It is essential that the Overviews be read for both sections. They explain the basic rationale of the sections and the nature and interrelationship of chapters.

Because the text is an introductory and general one, the treatment is limited largely to things of instructional significance. Such matters of administrative importance as teacher rating, curriculum evaluation, school plant appraisal, and community surveys are omitted. Moreover, the construction of standardized tests, the administration of individual intelligence tests, projective techniques for personality measurement, and other very special psycho-

metric procedures are discussed only with a view to general understanding and not to practice.

There are several exercises in each chapter designed to help you apply the principles and procedures discussed. In many cases you may wish to select from among them rather than do them all.

The bibliography at the end of each chapter indicates the readings that have contributed to the ideas and techniques presented in the chapter, including specific references. The bibliography is not meant to be a list of suggested readings. Where additional reading seems advisable, appropriate titles are indicated in the text.

Footnotes are used for technical explanations, for suggested additional reading, for certain citations with no general relationship to educational measurement or evaluation, and for passages that may interest only a few readers.

The appendix contains an extensive glossary of terms that have technical significance either in this text or in measurement and evaluation generally. While we have tried to define any unusual term the first time it is mentioned, and very important terms repeatedly, you may wish to consult the glossary from time to time. Also in the appendix are sample report cards and an annotated bibliography of published tests in all areas. The bibliography should be helpful in selecting tests for study and for use.

The appendix is concluded by two statistical tables. The table of normal curve area- z score relationships will be helpful in interpreting the confidence limits of various measurements. The other table compares graphically the several types of norm scores used in standardized tests.

ACKNOWLEDGMENTS

Textbooks are written only with the assistance and cooperation of a great many people. Here we wish to mention a few of those to whom we are particularly indebted.

Dr. H. Glenn Ludlow, Associate Professor of Education, University of Michigan, read the entire manuscript and provided numerous suggestions for its improvement. Dr. H. Orville Nordberg, Chairman of the Division of Teacher Education at Sacramento State College, read several chapters critically. Many others of our colleagues at Sacramento State College assisted us with encouragement, advice, and materials. Among them are Dean Harold B. Roberts, Dr. Palmer A. Graver, Dr. Edwin L. Klingelhofer, Mrs. Lucille Colby, and Dr. Charles F. Howard.

Three stenographers typed the bulk of the manuscript, Miss Gerry Millard, Mrs. Louise Lowry, and Mrs. Carolyn Larsen. Mrs. Larsen, who worked on the final draft, was as valuable for her editing as for her typing. We are extremely grateful to Mrs. Frances Jones, a young artist, who drew the many graphs and charts in the book.

The specimens of student work to be found in Chapter V were provided by the following teachers at Sacramento Senior High School: Mr. Frederick W. Blodgett, Mr. Clarence I. Fiscus, Mrs. Helen Lafferty, Mr. Ray E. Loehr, Mr. John P. Moore, Dr. Gerald G. Smith, and Mr. Gerry S. Watt. Sample report cards were sent to us by numerous elementary and secondary schools in the Sacramento region.

Many publishers have permitted us to reproduce items from their publications and, as in any textbook, we have drawn on and referred to many professional books and articles. These are all acknowledged as they occur.

Finally, we wish to thank the several hundred students who were in our educational measurement classes during the last five years. They were the "guinea pigs" for the development of the book. On them we tried out new ideas and new approaches, and from them we hope we learned something of what an introductory text in educational measurement should say and how it should be said.

JAMES M. BRADFIELD
H. STEWART MOREDOCK

Sacramento State College

CONTENTS

I. Fundamental Conceptions and Procedures: Overview	1
1. FORMS OF MEASUREMENT SYMBOLS	5
2. PREPARING PHENOMENA FOR MEASUREMENT	17
3. PROCEDURES OF MEASUREMENT IN GENERAL	34
4. OBSERVATION	48
5. PRODUCT ANALYSIS AND FREE-RESPONSE PROCEDURES	63
6. GUIDED RESPONSE PROCEDURES	84
7. STATISTICAL DESCRIPTIONS OF MEASUREMENT DATA	130
8. FURTHER STATISTICAL CONCEPTS IN MEASUREMENT	163
9. EVALUATIVE STANDARDS—MARKING AND REPORTING ACHIEVEMENT	190
 II. Customary Uses of Measurement and Evaluation in Education:	
Overview	215
10. LANGUAGE ARTS	219
11. SOCIAL STUDIES	266
12. SCIENCE AND MATHEMATICS	294
13. PERFORMANCE—ACTIVITY AREAS	329
14. INTELLIGENCE	361
15. PERSONALITY AND CHARACTER	386
16. SCHOOL-WIDE TESTING PROGRAMS	424
 Appendix	
A. GLOSSARY OF TERMS USED IN MEASUREMENT AND EVALUATION	441
B. BIBLIOGRAPHY OF SELECTED PUBLISHED TESTS	462
C. SAMPLES OF REPORT CARDS AND RECORD FORMS	494
D. NORMAL PROBABILITY AND DERIVED SCORE TABLES	498
 Index	501

TABLES

1. Some Pupil Products Particularly Amenable to Factor Count Scoring	79
2. Types of Guided Response Items Compared as to Chance Factor, Time for Administration, and Difficulty	102
3. An Outline for Constructing a Frequency Table	133
4. An Outline for Constructing a Histogram	135
5. An Outline for Computing the Median of a Distribution	143
6. An Outline for Computing the Mean from the Formula $X = X_o + \frac{i\Sigma fu}{N}$	148
7. An Outline for the Computation of the Standard Deviation	148
8. An Outline for Developing a Cumulative Percentage Curve	154
9. Published Tests Listed for Language Arts Areas in the <i>Fourth Mental Measurements Yearbook</i>	221
10. An Illustrative Plan of Measurement and Evaluation for Eighth-Grade Social Studies	282
11. Test of Attitudes Toward Government, Minority Groups, and School Life	285
12. A Free-Response Test Scored According to an Analysis Schedule	286
13. Observation Schedule for Study Skills and Attitudes	287
14. Pupil Study Log	288
15. Example of an Evaluative Standard for Science	300
16. Examples of Procedures for Measuring Ability to Think Scientifically	302
17. Example of an Evaluative Standard for Arithmetic Computation	314
18. Example of an Evaluative Standard for Understanding an Arithmetic Concept	315
19. Example of an Evaluative Standard for Logic	318
20. Examples of Test Items for Arithmetic	320
21. Examples of Test Items for Algebra	321
22. Examples of Test Items for Geometry	322
23. Rating Chart for Art Performance	345
24. Check List for Driving Performance	346
25. Musical Performance Test	347

26. Rating Form for Applying Varnish	349
27. Plan for Judging a Typing Performance	349
28. Outline for Rating a Gathered Skirt	350
29. Plan for Evaluating a Diving Performance	351
30. A Picture Articulation Test for Nonreaders	352
31. Outline of Personality and Character Dimensions, Measuring Procedures, and Evaluative Standards	390
32. Percentage of Area Under the Normal Curve Between Mean Ordinate and Ordinate at Given Z Score	498
33. Several Types of Score as Related to a Normal Probability Distribution	499

FIGURES

1. Principles of Sampling	39
2. Probable Relationship Between the Reliability or Validity of a Procedure and the Time and Energy Devoted to it	44
3. Check List for Observation in Physical Education	52
4. Anecdotal Record for Physical Education	53
5. Types of Rating Scale Items	56
6. Examples of Pupil Products Scored by Various Methods	64
7. Examples of Free-Response Items Scored by Several Methods	66
8. Sample Pictures from the Murray Thematic Apperception Test and the Rorschach Ink Blot Test	71
9. Examples of Guided Response Test Items Which Require Selection of an Answer	86
10. Examples of Guided Response Items Which Require Provision of an Answer	87
11. Examples of Guided Response Items in Which Arrangement of Elements is Involved	88
12. Measurement of Many-Degreed Dimensions by a Series of Dichotomous Classifications	95
13. Section of an Item Analysis Table	99
14. Illustration of a Test Item Entered on a Card Together with an Analysis of Responses to the Item	100
15. Example of a Strip Key and a Stencil Key	117
16. A Test Profile Form	125
17. Frequency Polygon	137
18. Smoothed Frequency Curve	137
19. Bell-Shaped or Normal Distribution	138
20. Symmetrical Distribution	138
21. Skewed Distribution	138
22. J-Shaped or Poisson Distribution	138
23. Rectangular Distribution	138
24. U-Shaped Distribution	138
25. Bimodal Distribution	138
26. Median Defined in Terms of the Area of a Histogram	141
27. Finding a Median Illustrated	141

28. Locating the Median in the Middle Interval	142
29. Mean as a Point of Balance	145
30. Illustrating the Mean as the Center of Balance of the Histogram	149
31. The Effect of Skewed Distributions on the Mean and Median	149
32. Two Distributions Having the Same Range but Different Variability	151
33. Probabilities in Coin Tossing	164
34. Normal Curve and Z Scale Relationship	166
35. Portions of Area of Normal Distribution Contained Between Given Z Scores	166
36. Probability of a Person's Having a Given IQ	168
37. Standard Error of an IQ Illustrated	170
38. A Test Profile that Includes an Indication of Standard Errors of the Scores	172
39. Illustrative Scattergrams for Different Degrees of Correlation	176
40. Sample Items from Readiness Tests	226
41. Types of Guided Response Items Used to Measure Various Dimensions of Reading Ability	231
42. Grading Sheet from the <i>College Entrance Board General Composition Test, Form F</i>	241
43. Examples of Guided Response Items Designed to Test a Pupil's Knowledge of the Mechanics of Composition	243
44. Examples of Guided Response Items and Techniques for Measuring Spelling Ability	245
45. Two Examples of Guided Response Items Keyed to Rhetorical Skill	247
46. How Evaluative Standards Seem to Operate for English Composition	249
47. Samples from the <i>Ayres Handwriting Scale</i>	252
48. Examples of Guided Response Items for Measuring Literary Knowledge	256
49. A Profile Reporting Form for English or Language Arts	261
50. Two Usual Types of Guided Response Items Used in Science	301
51. Profile Page of California Test of Mental Maturity	373
52. Illustrative Guess-Who Items Used to Measure the Reputation Pupils have with their Peers	400
53. Simplified Illustration of the Compilation of a Sociogram	401
54. Illustrative Sociogram of Friendship Patterns	402
55. Example of a Form to Report on Citizenship Items	404
56. Example of a Legal Evaluative Standard	406
57. Examples of Guided Response Items Used in Getting Pupils to Report on Their Own Interests and Attitudes	409
58. Examples of Guided Response Items Used to Measure Pupil Opinion	411
59. Examples of Three Types of Devices Used in Rating Personality-Character Variables after Observation	417

SECTION I

FUNDAMENTAL CONCEPTIONS AND PROCEDURES

OVERVIEW

In this first section of the book we wish to discuss certain conceptions and procedures that are basic to measuring and evaluating pupil achievement and other things of importance in the schools. As you will find, all the many principles and processes described are interrelated and pertain to a given definition of measurement and evaluation. In our overview of the first section we wish briefly to state this definition and then to explain the order and significance of the nine chapters in the section.

By now, all of us have some idea about the nature of measurement, particularly the measurement of our physical environment. How many times have we measured temperature by reading a thermometer, measured length by using a ruler, measured the speed of a car by reading a speedometer, and measured time by looking at a clock? Moreover, we have had countless experiences with evaluation in our everyday activities: judgments about this dress or suit as compared with that one, decisions as to party and candidate in elections, and the continual evaluations we make of our acquaintances. Then, too, in our years of schooling we have been subjected to innumerable measures of progress in school subjects, and each quarter or semester we have had our achievement evaluated by our teachers in terms of *A*'s, *B*'s, and *F*'s. So we should have some notions about the measurement and evaluation of educational things as well.

With all this background of experience concerning measurement and evaluation, each of us should be able to define them. So let us try it now. Write down your definitions of measurement and of evaluation. Then compare your definitions with those of your fellow students. Notice how they vary, how some are generalized and others particularized, how some omit what seems important to you and how others seem to include nonessentials. From our observation of the definitions submitted by hundreds of college students at the begin-

ning of a course in educational measurement, we will guess that your definitions vary so greatly that you well may wonder what you are going to study.

Consequently, in the interest of clear communication, we ask you now to abandon your personal definitions and accept the one we offer. Like any definition, ours is arbitrary but it is designed to cover the activities that we customarily call measurement and evaluation and it is further designed to make their study a rational and systematic process. We think you will find that the definition covers yours but may go beyond it or may require some change in your perspective.

Definition of Measurement and Evaluation

1. Measurement is the process of assigning symbols to dimensions of phenomena in order to characterize the *status* of a phenomenon as precisely as possible.

2. Evaluation is the assignment of symbols to phenomena in order to characterize the *worth* or *value* of a phenomenon, usually with reference to some social, cultural, or scientific standard.

To illustrate this definition, let us consider the case of a student in a tenth-grade typing class. Along with other students, he was given a five-minute *speed test* and was found to type *fifty words per minute* with a total of *seven errors*. This is the process of *measurement*. The *phenomenon* in question was typing. The *dimensions* to be measured were speed and accuracy. The *status* of the boy's typing relative to speed and accuracy was precisely *characterized* by the *symbols*, fifty words per minute and seven errors.

Next, the instructor consulted a *manual* that gave the degree of speed and accuracy to be *expected* of students in the tenth grade after given amounts of instruction and accordingly assigned the student's test a grade of *B*. This is the process of *evaluation*. The *symbol B* characterized the *worth* or *value* of the student's speed and accuracy in typing. The *B*, meaning "good," was assigned because the boy's speed and accuracy exceeded to a certain degree what was to be expected, the *standard*.

Plan of the Section

The first nine chapters of this book are essentially an expansion of this definition of measurement and evaluation, together with its corollaries. The symbols used in measurement are first discussed in Chapter 1. There we will find that measurement symbols are of three basic types: those indicating scale position, those indicating rank position, and those that simply denote a classification or constitute a description. The characteristics of measurable dimensions are discussed in Chapter 2, together with how dimensions may be construed in order to make them more measurable.

In our definition of measurement we stated that measurement is a process. This implies that certain procedures are used, and, in Chapters 3, 4, 5, and 6, these procedures are discussed as they relate to behavioral measurement. There

we find that measurement symbols may be assigned to dimensions either directly by the observer or indirectly through means of some instrument, technique, or device. The construction of these instruments and devices is discussed in detail.

Chapters 7 and 8 develop certain statistical concepts and operations that are important for the measurement process. Finally, in Chapter 9 the activity of evaluation is analyzed, together with the problems of marking and reporting grades. In this chapter, a great deal of attention is given to the development and use of evaluative standards.

In expanding the basic definition, the nine chapters not only describe how measurement and evaluation are performed but they also establish the variables that affect their validity and efficiency. These are:

1. The nature and appropriateness of the symbolic forms used to characterize dimensions (Chapter 1).
2. The relative "measurability" of the dimensions appraised and the suitability of these dimensions to the nature of the phenomenon and the purpose of measurement (Chapter 2).
3. The aptness of the procedures used to achieve the assignment of symbols to phenomena (Chapters 3, 4, 5, and 6).
4. The appropriate application of statistical concepts and use of statistical techniques to recast the measurement data into interpretative form (Chapters 7, 8).
5. The appropriateness and validity of standards selected for evaluation and the efficiency with which status is compared with standard (Chapter 9).

This, then, is a prospectus of the first section. It can also serve as a summary and might well be reviewed when the section is concluded.

CHAPTER I

FORMS OF MEASUREMENT SYMBOLS

The end result of measurement is the most familiar and obvious of its several aspects. This is the symbol or symbolic expression used to characterize the status of something. All around us, in newspapers, magazines, over the radio, the television, and in conversations, we see and hear numbers and other symbols expressing the results of measurements: 60-foot frontage, 1240 kilocycles, 80 degrees, 25 miles per hour, 125 pounds, 85 cents per dozen, 3:30 P.M., and 2.3 decibels. We are more familiar with the symbols of measurement than any other aspect of the process simply because they are what is communicated. After a scale has been used or a test applied, we are no longer concerned with it, but only with the measurement it accomplished. It is this symbol that we write down and talk about and use.

In this chapter we shall analyze the symbolic expressions of measurement, not only in order to understand them but also in order to gain some initial insight into the measurement process itself. First, we shall illustrate the several types of symbolic expressions used by man in his efforts at measurement. Following this, each form will be subjected to more careful scrutiny and compared with the others. Finally, the different forms of measurement will be viewed as a whole in terms of their significance for educational measurement.

Types of Measurement Symbols in General

We have no precise historical record of the first forms of measurement used by man. We can, however, speculate that one of the earliest attempts at measurement made by our ancestors was in connection with determining "how many." This problem must have arisen early in the cultural development of man—just as soon as there was concern about how many sheep or cattle were possessed, how many warriors were in a given tribe, how many days had passed, and how many fish to catch in order that no one in the group would go hungry.

According to archaeological findings, we have good reason to believe that the earliest expressions of "how many" were symbols meaning "few," "many," and possibly symbols for "one" and "two." Certainly, in some of the primitive cultures existing today, the question of "how many" is answered only by terms that mean "one," "two," and "many." At the same time or possibly a little later, man developed symbolic equivalents of "more than" and "less than" in order

to express his observation that some collections contain more objects or fewer objects than other collections.

Later in the development of man's conception of number, fingers, pebbles, and notches in wood were used as symbols for indicating the quantity of objects in a collection. Tally marks soon became systematized into primitive number systems, and finally, today we have the Hindu-Arabic number system that can represent "how many" in any situation and to any desired degree.

This sketch of how man has devised measures for one important dimension, quantity, serves to illustrate the basic forms of measurement symbols he has developed and continues to use.

1. *Symbols that classify or describe.* The symbols "few" and "many" are of this type. They indicate simple categories or classes of quantity, and collections are characterized by the categorical term assigned them.

2. *Symbols that indicate rank or order.* "More than," "less than," "largest," "next largest," and "smallest" are examples of measurement symbols that indicate rank. Here collections have been compared and designated in terms of their differences in quantity.

3. *Symbols that indicate a scale position.* The symbols of this type used in measuring "how many" are the numbers of the Hindu-Arabic system such as "315," "27," "9." The numbers characterize collections as to their quantity by stating where they belong on a counting scale.

It is not necessary to rely on history to see the use of these three types of measurement symbols. For instance, we might observe a child's growing conception of temperature. His first measurements are usually "hot" and "cold." Later on he may add further classification symbols such as "warm" and "cool." Soon he is able also to rank objects on the basis of their temperature by using such rank or comparative symbols as "hottest," "warmer than," and "coldest." The final stage is reached when he is able to report the temperature of an object by use of a symbol that indicates a position on the temperature scale, e.g., 87 degrees or 62 degrees Fahrenheit.

Another example is provided by a child's development of expressions for speed. At first it is simply "fast" and "slow"—measurement symbols that classify or describe. Then follow the symbols "fastest," "slower than," and "slowest"—measurement symbols that indicate comparisons of rank or order. Finally, the child can express with understanding the speed of an object in miles per hour—measurement symbols that indicate a scale position.

While numerous other examples could be provided, both from the developing concepts of children and from science as a new phenomenon is studied, these should suffice to illustrate the three basic forms of measurement. Now we need to examine each type in detail.

Measurement Symbols That Express a Precise Scale Position

We shall discuss first the symbols that express a definite scale position, since this form is most commonly associated with the word "measure." The

examples provided at the beginning of this chapter—60-foot frontage, 125 pounds, 85 cents, 80 degrees—are all symbols that express a scale position. Each time we read a ruler, a clock, a speedometer, a weight scale, a thermometer, or count objects, we are finding a measurement symbol that expresses a scale position.

A scale consists of a series of units, each of which represents an exactly defined portion of or degree of variation in the dimension being measured. It is, of course, preferable to have the unit universally standardized so that it may be more widely used. In expressing "quantity" by an Arabic numeral, the unit is considered to be the individual indivisible object that comprises the collection being measured. In the case of "how many people," the unit is a "person" and in the case of "how many years," the "year" is considered the unit. We also know that a unit of weight is the pound, a unit of length is the inch, and so on. In the expressions of scale positions, we always find the unit included or at least inferred—"3 inches," "3 dollars," "3 seconds," "3 pounds." Ordinarily, the unit is constant throughout the scale and, if this is the case, then the difference between any two adjacent scale positions is everywhere the same. This idea is illustrated by the fact that the difference between a length of 78 inches and a length of 75 inches is the same as the difference between a length of 11 inches and a length of 8 inches.

Most of the scales that we have seen in our everyday affairs have a zero point, which ordinarily represents complete absence of the phenomenon being measured. On our scale of "quantity," zero would represent a collection containing no objects. Also, on a weight scale, zero would represent the absence of any weight, and, on a tape measure, zero would represent the absence of length. Measurement symbols that refer to positions on scales that have zero points can be added, subtracted, multiplied, and divided, and, furthermore, ratios can be computed. For example, if an object weighing 28 pounds were combined with an object weighing 14 pounds, the weight of both objects together would be 42 pounds. In this case, the measurement symbols 28 and 14 can be added meaningfully and, furthermore, these two symbols can be subtracted, multiplied, and divided, and the results will have meaning as far as this particular scale is concerned. In addition, these two symbols can be set into ratio with each other by stating that the second object weighs half as much as the first object.

Several scales, although they have definitely established units, do not have absolute zero points. The ordinary Fahrenheit and centigrade temperature scales are examples of this type of scale. While each of these does have something called a zero, these points are purely arbitrary and do not represent the point of complete absence of temperature that is called absolute zero. Another example is the scale of IQ's, a complete absence of intelligence having yet to be found. Measurement symbols that refer to scales without absolute zeros still can be added, subtracted, multiplied, and divided, but ratios should not be computed. We do not say that 80 degrees Fahrenheit is twice as hot as 40

degrees Fahrenheit, nor do we think that a person with an IQ of 120 is twice as intelligent as a person with an IQ of 60.

The crucial consideration for use of a scale form of measurement is, of course, the existence of defined portions or of defined units of variation in the dimension in question. When these are not present, scale measurement is inapplicable, despite its obvious advantages. Unfortunately, many of the dimensions of educational phenomena have no such defined units and hence may not be measured by scale symbols, but instead must be appraised by the forms that require no definite units, i.e., ranking, classification, and description.

Measurement Symbols That Express a Rank or Order Position

To explain rank or order symbols, let us analyze a situation in which this form of measurement normally is used. Suppose a sales manager wishes to indicate the sales ability of one of the salesmen on his staff of fifteen salesmen. He could say simply that he is the best salesman, or the worst, or near the top. These terms are gross indicators of rank and are like the primitive and childish expressions we discussed earlier. A more precise measure would require that he arrange the fifteen salesmen in the order of their sales ability and then note what numerical rank falls to the salesman in question. It would be possible for the salesman to have a rank position ranging from 1, which is the top position, down to 15, which is the bottom position. The number of his rank would be the measure of the man's sales ability relative to this particular group of salesmen. Rank is useful as a measurement symbol in many other situations: beauty contests, the judging of animals and exhibits at a fair, and, of course, in school. It does not require units and a separate scale but only comparisons among individuals or objects.

While in our example Arabic numbers were used as symbols of rank or order, any set of ordered symbols can be used for this purpose: *A, B, C, D*, or $\alpha, \beta, \gamma, \delta$, or I, II, III, IV. It is important to notice that these symbols have little meaning when they stand alone. For example, if we were told that a certain person had a rank position of 13 with regard to sales ability, we would know very little unless we were also told the size of group in which the ranking took place. A rank of 13 would be relatively high in a group of 432 but relatively low in a group of 15.

In order to overcome this deficiency of rank numbers, percentile ranks have been developed. These indicate the percentage of any group that lies below the person in question. Hence, to assign a person a percentile rank of 78 would mean that he is better than 78 per cent of the group. This standard way of expressing relative position is discussed in detail in Chapter 7.

Symbols that indicate relative position usually are numbers and they attempt to serve the same basic measurement function as do the numbers that indicate a scale position. However, they involve basic ideas and procedures that are quite different from unit-scale numbers. Some of these differences are cited below.

1. *Rank symbols generally provide for only a limited amount of comparison.* One of the more obvious limitations of rank symbols is that they have meaning only in terms of the group in which the rank was established. If a student stands third highest in spelling in his class, this rank symbol has no meaning with respect to any other class or group of students. It is quite conceivable that the rank of this student's spelling achievement would change considerably if he were placed in another group. On the other hand, scale symbols, which are based on standardized units, can provide for universal comparison. A student's height symbolized in inches provides a basis for making a comparison with the height of anyone else in the world. A rank symbol does not have this possibility.

Steps can be taken, however, to offset this limitation of rank symbols. The most obvious step is to extend as much as possible the range of the group in which the ranking is done. Another step is to rank on the basis of a group that is representative of a much larger group. Rank symbols based upon such a standard group are called *norms* (see Chapter 7).

2. *Rank symbols cannot indicate the extent of difference between adjacent ranks.* If a student ranks fifteenth in his class in arithmetic achievement, there is no indication of how much better he is than the student whose rank is sixteen, or less than the student with rank fourteen. On the other hand, with scale symbols we can determine precisely the difference between two students.

3. *Rank symbols are limited as to what can be done to them mathematically.* In connection with scale symbols, we noted that they can be added, subtracted, multiplied, and divided meaningfully. Rank symbols cannot be so manipulated. Adding a rank of 3 to a rank of 5 would certainly be a meaningless operation. In Chapter 7 we shall indicate what mathematical operations can be performed with rank symbols.

4. *Scale symbols may be converted into rank symbols but rank symbols may not be changed to scale symbols.* If we know a student's height in inches, for example, we can easily find his rank in height among the other members of his class. We need only to have the height in inches of each member of the class and then arrange these in order of size and note the relative position of the student in question. On the other hand, rank symbols cannot be changed into scale symbols for the reason stated in 2 above.

All these specific differences between unit-scale symbols and rank symbols stem from one basic difference. Unit-scale symbols are based upon a unit of measure, generally a standardized unit, while rank symbols of measurement are *not* based upon any unit of measure but only upon an observed difference between individuals or objects. The amount of this difference is unimportant for determination of rank, but it is exact and critical for scale measurement.

Measurement Symbols That Classify and Describe

The third form of measurement used by society again does not make use of a unit of measure and yet it is extremely valuable in characterizing the status

of many phenomena. It involves simply an indication of the classification or category to which something belongs. Model numbers are an example. By saying that a certain car is a 1952 model, we indicate that it belongs to the class of cars made in 1952. Similarly, we can speak of a B-47 airplane, a model 3C242 centrifugal pump, and a model S62 office desk, all of which symbols indicate the category to which the objects belong. Other examples of measurement by classification are the symbols used in Selective Service, serial numbers and file numbers, the Dewey decimals on library books, and various job classifications, e.g., GS-5 in the federal civil service.

Each classification symbol ordinarily represents a description, drawing, picture, or set of specifications that portrays the distinguishing characteristics of the persons or objects comprising the classification. In Selective Service classification, II-A means:

In class II-A shall be placed any registrant found to be a "necessary man" in any industry, business, employment, agricultural pursuit, governmental service, or any other service or endeavor, or in training or preparation therefor, the maintenance of which is necessary to the national health, safety, or interest in the sense that it is useful or productive and contributes to the employment or well-being of the community or the nation.¹

In the same manner, the classification symbol "1952 model" for a particular automobile represents a picture or outline drawing of the car and a description of such characteristics as the over-all dimensions, number of cylinders, type of carburetion, shock absorption, engine mounting, water cooling, brake action, and grillwork.

Often, instead of letters or numbers, words or short phrases may be used as classification symbols. A good example is found in the words often used by the Weather Bureau to express the status of wind. These are: calm, light breeze, gentle breeze, moderate breeze, fresh breeze, strong wind, gale, and hurricane. Each term symbolizes a descriptive category. For example, "moderate breeze" stands for the following description: "raises dust and loose paper, small branches of trees are moved and the leaves and small twigs are in constant motion, extends a heavy flag."

As a further example, consider a teacher who wishes to appraise pupils with regard to acceptance of authority. He could use, and many teachers have used, words such as these to represent categories of variation regarding this dimension: defiant, sullenly complaint, ordinarily obedient, respectful, and completely docile. Each of these classifications needs to be described in detail so that the symbols will have a specific meaning. What the teacher uses in this case is generally called a behavior-rating scale. This device is discussed in detail in Chapter 4, pages 54-59.

From our discussion, it is apparent that the role of the class symbols them-

¹ Selective Service Regulations, Volume III, *Classification and Selection* (Washington, D.C.: U.S. Government Printing Office, 1940), p. 19.

selves is purely nominal—they simply serve as convenient shorthand expressions. The crucial part of this form of measurement is the description of the various categories. For effective classification, the description should spell out as completely as possible the distinguishing characteristics of the category. This should include, if necessary, the typical situations in which these characteristics are observed. Secondly, the description should clearly mark the boundaries between adjacent categories. Ordinarily, this is done by drawing contrasts between items belonging to different categories. These requirements for the descriptions suggest that setting up a scheme of classifications for a given phenomenon is not an easy task.

It is important to note here that a classification scheme can deal with but one dimension or aspect of a person or object, and that it provides only for gross characterization as to this one dimension. The Selective Service classification system, for example, is focused specifically upon eligibility for service in the armed forces and it contains only a few categories of eligibility. As a result, men who are quite different in many respects may find themselves in the same category. An electrical engineer, a physician, and a dentist might all be given a II-A classification. Moreover, each II-A might actually deserve slightly different treatment by a draft board. Similarly, the classification scheme for acceptance of authority, which was provided earlier, ignores differences in honesty, intelligence, age, etc., and also the slight differences in acceptance of authority that would necessarily obtain for all pupils classified as "ordinarily obedient."

Teachers often need a broader and more precise characterization of their students than classification provides. This occurs particularly when the teacher wishes to evaluate such things as study habits, citizenship, social adjustment, etc. To classify pupils regarding these complex dimensions would simply not help much. So the teacher must resort to *describing* the pupil with many words, phrases, and sentences.

Description and classification are, of course, hardly separable. If we were to consider that each individual in a group *could* constitute a category in himself, then the description of the individual would be the description of the category and the forms would be synonymous. In practice, though, we think of classifying when the categories are much less numerous than the individuals (as with selective service). Moreover, the words and numbers that symbolize classifications are not essentially different from the words and numbers that are used to describe. But again in practice we think of certain words more as classifications (genius or moron) and others more as simple describers (highly intelligent or very dull).

In description, the intent is primarily to identify, to distinguish, to characterize an individual informally with respect to an unlimited number of aspects or dimensions. In classifying, on the other hand, and in scaling and ranking, we are more concerned with comparisons and appraisals with respect to a single dimension or aspect. What is necessary for a valid category description

is likewise necessary for a valid description of an individual. It must indicate the salient characteristics and the boundaries of the phenomenon. Words, letters, and numbers are all used in description as they are in classification, and frequently some of the descriptive entries are expressions of rank, scale position, or classification for components of the phenomenon. The form does not facilitate comparisons among individuals, but it can characterize the individual's unique status as no other form can.

Description is widely used by teachers and psychologists, both for the reason of its inherent validity for complex phenomena and because other forms are too often inapplicable. Application of the form is discussed in detail in Chapter 4. Here it may suffice to offer an example of a description as used in education. The following is quoted from a record of an interview made by one of the authors:

... drew a picture of his family and discussed his feelings with the examiner. The picture figures were stereotyped and more like those of an eight-year-old than a twelve-year-old. They disclosed no particular family feelings or relationships. His very rapid speech and his occasional stuttering both became accentuated when he was questioned about his family. He answered questions about them readily and his answers connoted defensive pride to the examiner. He volunteered that he used to steal but stopped when his father whipped him. He thinks stealing is wrong. When questioned further about what he had stolen and when, he flushed and became incoherent in his relation. No details were elicited. He admitted that his twin still steals.

Forms of Measurement Viewed as a Whole

Up to this point we have discussed the various forms of measurement symbols and the advantages and difficulties of each of them. Numbers, words, and letters, singly and in groups, measure as they indicate scale position, as they designate rank or order, and as they classify or simply describe. We have treated the forms separately, though contrasting them, and in so doing we have oversimplified the matter. We have acted as though a phenomenon to be measured might, as a rule, be measured as a whole and by a single form. This, of course, is seldom the case. Most phenomena of interest to educators have varied aspects that require different forms of measurement. Consequently, to measure these phenomena requires a combination of forms.

To illustrate, consider a boy's "academic aptitude." A very minimum measurement of this phenomenon involves some scale symbols: age twelve, IQ, 110, vision in both eyes 20/20. It entails some rank or order symbols: 80th percentile in arithmetic in the 7th grade, has a reading age of 12-2. It necessitates some classification: he is a *B* pupil and a "good citizen." And, finally, only a description of his study habits may be given.

A choice among the several forms—scale, rank, classification, and description—is more a function of the phenomenon to be measured and the purpose of the measurement than it is of the forms themselves. Of course, it

pays to measure with numbers because arithmetic may be applied to them. It is better to scale than to rank, because scale numbers can be added, subtracted, etc., while rank designates cannot be. And if comparison among comparable phenomena is desired, description is of little help. But, beyond these considerations, you should determine the forms to be used on the basis of what you wish to measure, for what purpose you measure, and the evaluative standards to be applied. The details of this interrelationship between form, object, purpose, and evaluative standard are specified in the several chapters of Section II.

Whatever symbolic forms are used, they need to meet certain conditions if accurate measurement is to be accomplished. These are in addition to, or perhaps in re-emphasis of, some of the conditions specified for the particular forms.

1. *The forms should be appropriate to the type of variation to be measured.* As we will see in the next chapter, educational phenomena manifest two general types of variation. Some, such as intelligence, involve serial variation. Pupils differ as to degree or amount and, consequently, either scaling or ranking may be used because both express differences in degree or amount. The second basic sort of variation is that of type or category. This is exemplified by vocational and recreational interests. Here pupils differ as to the type of work they like to do and the type of sport or game they prefer. Obviously, some sort of classification or descriptive form must be used to characterize such categorical variation.

2. *They should exceed the range of possible variation in the thing being measured.* Without this, cases touching or exceeding the limits of the scheme are unmeasured. An example is a ruler painted on the wall, starting at four feet and running to seven. Some midgets, less than four feet high, and some giants, more than seven feet tall, are of unknown height according to this scale.

3. *They should cover the range of variation completely.* Obviously, if a gap is left, we do not measure with any precision a case that happens to fall within the gap. This error is frequently made in dichotomous classification. For example, if in a sociological survey we try to classify all persons as married or unmarried, we will have difficulty with "common law marriages" and even more casual relationships in which children and parenthood are involved.

4. *The symbols should have standard and limited meaning.* Since most, if not all, measurement involves or results in communication, it almost goes without saying that our terms must mean essentially the same thing to all who are to use them. Moreover, if we do not know where one symbol, say *upper-middle class*, ends and another, *lower-middle class*, begins, we have a great deal of trouble appraising those cases that fall near what should be the limit of the symbol's meaning.

5. *Finally, the appropriate form of measurement should be selected without regard to preconceptions about the superior value of numbers.* Unfortunately for education, numerical measurement has a tremendous amount of

prestige. This is primarily because the nearly miraculous progress of the physical sciences has been associated with the use of numerical measurements. Many of us who practice education have come to believe that effective measurement simply has not been accomplished unless numbers have been obtained. Moreover, we are apt to get a certain feeling of security in dealing with numbers. They are exact and objective, and furthermore pupils are not prone to question numerical results.

Because of our identification of numerical measurement with scientific measurement, some teachers, administrators, and even psychometrists and psychologists have made serious errors in measuring pupil achievement. Too often we have appraised only the more superficial aspects of pupil knowledge and skill simply because they can be measured by numbers. We have counted the number of pages a pupil has read, the number of homework problems he has completed, the number of punctuation errors in his sentences, the number of recitations he has made, and, of course, the number of true-false questions he has answered correctly. This we have done, to the neglect of his understanding of what he has read, what he has learned from his homework, the smoothness of his sentences, the quality of his recitations, and the merit of his reasons for answering true or false. The latter category of things demonstrably is more significant than the former, but it is more difficult to measure with numbers, and particularly with scale numbers.

We take the view that *effective* measurement is *scientific* and deserves prestige, not just a scale form of measurement. As a matter of fact, the "scientist" uses each of the other forms when the situation requires it. The chemist *describes* the attributes of a new synthetic fabric. The physicist *classifies* a type of radiation as beta or gamma. The engineer *ranks* building materials according to their durability. The difference between the measurements of the skilled teacher and the scientist is one of degree and not of kind. Because the objects of the teacher's measurements are the more complex and intangible, his measurements are the less precise. But when the scientist's objectives are equally complex and intangible, his measures are equally imprecise.

We should consider that form of measurement to be optimum which most precisely characterizes the essential dimensions of a phenomenon. In education, this means that classificatory and descriptive words are often the optimum forms because many of the dimensions we need to measure may not be validly characterized by raw numbers alone.

Summary

The purpose of a measurement symbol is to characterize the status of that which is measured. Forms in current use may designate scale position, rank or order within a group, classification, or simply constitute a description.

The most precise of these forms is scaling, and scale numbers may be manipulated mathematically. However, scaling requires a zero or other fixed reference point and the existence of defined and constant units of difference.

Since there are no constant units or zero points for most educational phenomena, scaling is seldom applicable. Rank and classification symbols may be used without reference to units or zeros and, consequently, are widely applicable in educational measurement. These two forms are less precise, though, and are not amenable to addition or multiplication. Description is the necessary recourse for many complex phenomena and, while descriptions are not easily compared, they can constitute very exact appraisals.

The primary considerations in selection of a measurement form are the objects of measurement and the standards by which they are to be evaluated. In addition, any system of measurement symbols should satisfy these conditions: (1) Be appropriate to the type of variation to be measured; (2) exceed the range of variation in the things being measured; (3) cover this range completely; and (4) have standard and limited meaning.

In the final analysis, the measurement symbols used in education should fulfill their intended function of characterizing the status of complex educational phenomena. Because of the nature of these phenomena, more reliance should be placed upon a combination of rank symbols and classificatory and descriptive expressions without regard to the superior "prestige" of numerical measures.

EXERCISES

1. Select a variable phenomenon such as sound, color, slope, or cost of living, and give examples of each of the three types of symbols that may be used in its measurement.

2. The statement is made: "The sky is blue." Is this measurement? Explain your answer.

3. If a phenomenon can be measured by a classification scheme only, what can be said about the phenomenon?

4. Identify the categories of a classification scheme for such phenomena as fair-mindedness, honesty, perseverance, study habits, muscular co-ordination, and social adjustment.

5. Susan has a vocabulary of 500 French words and Sally has a vocabulary of 700 French words. What type of measurement symbol is used? What can be said about the difference in French vocabulary between Susan and Sally? Explain your answers.

6. For each of the following measurement symbols often encountered in education, identify the type of symbol used and justify your answer:

IQ 105 on Stanford-Binet test

85 per cent on a teacher-made arithmetic test

Spelled 15 out of 20 words correctly

Strength of grip, 15 pounds

Received 113 votes in a school election

Mental age eight years, six months

Fifty-seventh percentile on a standardized test

Sixth-grade reading level

Received a *B* in the course

Grade point average or honor point ratio is 2.87

BIBLIOGRAPHY

1. Campbell, N. R., *An Account of the Principles of Measurement and Calculation*. London: Longmans, Green and Co., 1928.
2. Lorge, I., "The Fundamental Nature of Measurement," *Educational Measurement* (E. F. Linn, ed.). Washington, D.C.: American Council on Education, pp. 533-559.
3. Travers, R. M. W., *Educational Measurement*. New York: The Macmillan Co., 1955, Chap. III.
4. Stevens, S. S., "On the Theory of Scales of Measurement," *Science*, 103:677-680, 1946.

CHAPTER 2

PREPARING PHENOMENA FOR MEASUREMENT

In the chapter just completed we examined educational measurement from the viewpoint of the symbolic expressions we assign to things when we measure them. During this examination we discussed three symbolic forms, classification-description, ranking, and scaling. We discovered that they differ as to precision and significance and thus that the form of symbol to be used affects the adequacy of the measurement. So it is with the phenomena themselves, the subject of the present chapter. They differ widely and in their variation they affect the efficiency of the measurements made of them. Some are relatively easy to measure, others difficult. Some lend themselves to direct measurement, others only to indirect procedures. Some permit precise appraisals and many others, of course, are susceptible only to very gross measurement.

In our study of measurable phenomena we first shall review some of the problems posed by the array of educational things for which measurement may be desired. Then attention will be given to the way in which phenomena must be construed if they are to be measured. Several basic attributes of measurable dimensions will be cited. We shall discuss *inferred* dimensions and the dimensions of phenomena that constitute *constructs*. Some attention will be given to issues relevant to the selection of dimensions for measurement. Finally, we shall discuss the basic dimensions of pupil achievement in school subjects.

Educational Phenomena for Which Measurement May Be Desired

We use the word "phenomena" as a collective symbol for the possible objects of measurement because it is about the only word sufficiently general to encompass all the various things that teachers and other educational personnel say they wish to measure. The following list suggests the great variety of these phenomena: ¹

Ability (in)	Abstract thinking	Adjustment
Art	Achievement (in)	
Music	Arithmetic	
Tennis	Social studies	
Etc.	Etc.	Aggressiveness

¹ Derived from indexes of books on educational measurement and from the experience of the authors in teaching and in working with schools on measurement problems.

Aptitude (in)	Effort	Personality
Clerical tasks		
Engineering	Gain	Problem-solving ability
Physical education		
Etc.	Group structure	Readiness (for)
		Arithmetic
Attitudes, in general	Handwriting	Reading
and toward		
Democracy	Interests (in)	Skill (in)
Minority groups	Play	Cooking
School	Reading	Typing
Etc.	Vocations	Wood work
	Etc.	Etc.
Character		
	Intelligence	Speech
Citizenship		
	Knowledge (of)	Study habits
Community attitudes	Chemistry	
	Contemporary affairs	Understanding (of)
Creativeness	History	Geography
	Literature	History
Critical thinking	Etc.	Physical principles
		Etc.
Developmental age	Leadership	
		Vocabulary
Dexterity	Mental age	

The phenomena that educators wish to measure constitute an unsystematic array and they have several characteristics that make for difficulty in measurement. For one, the items are neither mutually exclusive nor collectively inclusive of all that may be significant. For example, skill, ability, and achievement each may refer essentially to the same thing, and so with aptitude and readiness, and with knowledge and understanding. Yet the uninitiated may think that they refer to different things and may try to measure them separately with foredoomed frustration. Then, wherever items are free of overlapping, they often do not cover all the ground they should. To predict a pupil's success in Grade IX science, we might think we need only to measure his intelligence, his previous knowledge, his attitude toward school, and his study habits, since these are the things usually measured in predictions of school success. Yet we could appraise these very precisely and still not predict success in Grade IX science with any great accuracy because there are additional things involved in school success: strength of motivation, efficiency of instruction, adequacy of materials, etc.

A second troublesome characteristic of the array is that too many of the items are high-order abstractions having no clear, agreed-upon definition.

Character, citizenship, knowledge, personality, and intelligence, for example, have nearly as many definitions as there are tests for them and the definitions too often are in terms of equally abstract words. We shall find in a few pages how important is clear definition for measurement and how difficult is the measurement of abstractions.

In the third place, measurement is used for a variety of educational purposes: marking, programing, prediction, diagnosis, research, administrative planning, curriculum evaluation, public relations, etc. Frequently a given purpose may involve a special point of view toward a phenomenon and, if several points of view exist for a single phenomenon, there may be confusion in its measurement. To illustrate this confusion, consider the measurement of reading. A teacher needs to give fifth graders a mark in reading; so she measures them to see how well each has learned to read the materials he has used. A junior high counselor needs to assign students to different English classes; so he measures the students to see how they vary in reading ability. A remedial teacher must help cases of reading retardation; so he measures to see just what are each pupil's reading difficulties. Finally, the superintendent wishes to impress the public with the excellence of the district's reading program. Consequently, he measures to see how far the mean² reading ability of each grade is above some norm³ for the grade. Because of their varied purposes, each of these persons is likely to use different tests of reading and to express the results of his measurements differently. Yet each may consider that he has measured the same thing as the other—reading.

Finally, most of the phenomena are behaviors, some are covert behaviors, and a few are terms for inferred states of mind or emotion that underlie behaviors. As such, they are processes, not things. They change; they occur and then they do not occur; often the act of measurement distorts them; and they have few to no exact physical properties. Consequently, the practice of repeated measurement under the same conditions that gives the physical sciences their exactitude is largely obviated in education by the very nature of the phenomena needing measurement.

Because of such characteristics, it is apparent that educational phenomena must be carefully examined with a view to their measurability. Many of them will need to be redefined or otherwise prepared before measurement is attempted.

The Nature of Measurable Dimensions

It is axiomatic that we have to measure things in terms of their aspects, properties, qualities, or dimensions. We say a man is 6 feet *tall*, is one 170 pounds in *weight*, and is a pinkish *color*. If we omit the dimension designation in our statement—just say that he is twenty-seven—we count on the listener

² For explanation of mean, see pages 144–149.

³ For explanation of norm, see pages 158–160.

to infer what we omitted, probably that his *age* is twenty-seven years. The names of these things relative to which we measure phenomena are many: factors, variables, properties, conditions, parameters, qualities, dimensions, etc. For convenience we shall use the single word *dimension*. Anything that we measure a phenomenon "in terms of" or "with respect to" we shall call a *dimension*.

If we measure phenomena only in terms of their dimensions, it follows that we can measure them only to the extent that we have identified their dimensions and only to the extent that these identified dimensions lend themselves to measurement. Viruses constitute a case in point. When viruses first were hypothesized by physicians as the causes of disease, they were hardly more than a word for "cause unknown," since only the property of "causing a symptom" or "not causing a symptom" was established for them. Appraisals of their status were conjectural, vague, and subjective. It was only as such definite dimensions as "size of filter through which they might not pass," "rate of growth," "preferred culture," "chemical make-up," etc., were indicated that physicians could say it was possible to measure viruses.

It should be noted that there may be only a relative distinction between a phenomenon on the one hand and a dimension on the other. Certain things may constitute the dimensions of a given thing; but when they are viewed just for themselves, they in turn have dimensions. For example, to determine the status of a school pupil we need to measure his age, height, weight, knowledge of school subjects, intelligence, personality, background, etc., such factors being basic dimensions of the pupil. But we may wish to measure his intelligence alone and to do so we identify and appraise its dimensions, such things as memory, perceptual discrimination, vocabulary, and reasoning ability.

Dimensions are considered to be measurable; that is, capable of being described, classified, ranked, or assigned a scale number, to the degree that they approximate five basic conditions. As will be apparent, some educational phenomena already possess identified dimensions that meet the conditions, but others do not. In the latter case it will be necessary to recast the dimensions so that they do meet the conditions. In some instances, entirely new dimensions may have to be designated.

1. *Measurable dimensions are common to a group or class of things.* First, as a matter of logic we should notice that a measurable dimension is relevant to a group and not to an individual only. Obviously, all the myriad pupil dimensions interesting to educators pertain to many if not to all pupils. Age, height, weight, reading ability, motivation, etc., are exhibited in varying degrees by all pupils. But suppose we had a dimension absolutely unique to one pupil, could we measure it? Suppose John was found to have a property that no one else ever had. If we are to measure it, we must assign a symbol to it that classifies or describes, ranks, or shows a scale position. Can we express John's status with respect to the property by classifying? Only by having two classes: pupils with the property and those without it. Can we use description?

That requires the use of descriptive words and/or numbers, and the very nature of words and numbers is that they refer to previously encountered properties or dimensions. Can we rank the boy with respect to the property? No, we have only one case. Finally, can we measure the thing in terms of a scale value? The existence of a scale requires the existence of previous phenomena from which the scale was derived, so we could have no scale. Apparently, then, we can appraise John's unique property only to the extent of saying that he has it and that no others do, which is no more than we knew when we started to try to measure it.

To reassert the condition, measurable dimensions are necessarily relevant to a group and not to one case or one individual only.

2. *A measurable dimension must provide sensory data.* A second important characteristic of measurability is demonstrated by the behavior of persons who engage in actions called measuring. A carpenter measures the length of a board by *looking*, to see what line on his jointed rule is at the end of the board. In judging the range of a pupil's voice, a music teacher *listens*, to determine which piano notes are like the highest and the lowest sounds the pupil can make. To measure the gap in a spark plug, the mechanic *feels* which of several thin leaves or wires will just fill the gap. We might go on to call on a cosmetician to appraise the strength of a perfume through *smelling* and a gourmet to determine the amount of garlic in a pot of soup through *tasting*. And even when the teacher measures his pupils' knowledge of history through the device of a true-false test, he finally *looks*, to see which questions are answered correctly and which incorrectly.

It is apparent from these examples that measurement entails some kind of sensory data; some person has ultimately to receive some sensation to accomplish the measurement.⁴ So, to be measurable, a dimension must provide some sensory data. Moreover, the more discriminating are the senses aroused, the more the dimension lends itself to direct measurement. For example, we *see* things more discriminately than we *feel* things. Thus, we measure height directly by *looking* at the pupil and the ruler, but we must measure weight indirectly through *looking* at a dial or a balance because the kinesthetic sensations by which we feel weight are too indiscriminate to provide a reliable measure. Finally, the more clearly differentiated are the sensations provided by the dimension, the more precise is likely to be its measurement. We are apt to be more accurate in counting the pupils in a given class than we are in judging how bright is the light in the classroom. The difference between six pupils and seven pupils is easy to see, whereas the difference between ten foot-candles of illumination and eleven foot-candles is less distinct.

⁴ In such devices as automatic pilots and gun directors, many measurements are made and are not observed by a person but, rather, by another part of the device. However, in these cases you may consider that a part of the device is an "observer," a substitute for a person, or that the pilot or gun chief *could* see, hear, or feel something if he wished to or that he could attach instruments whose movement he could observe.

Date... 15-3-85
Rec. No. 3064



In education, the problem of sensory data is particularly acute. Such things as speaking, singing, and studying can be observed but, being events, they ordinarily provide only nonrecurrent sensations. Such things as knowledge of history, understanding of science, intelligence, and attitudes toward school cannot be observed at all; so for them it is necessary to devise observable dimensions that constitute evidence of their *unobservable* dimensions.

3. *To be measured, a dimension must be clearly defined.* This third essential condition of measurable dimensions is self-evident. If to measure means to assign a symbol that properly characterizes the status of a phenomenon with respect to some dimension, it follows that the dimension must be clearly defined. Indeterminateness in the dimension necessarily means indeterminateness in status with respect to that dimension.

In education, the clear definition of dimensions is particularly critical since the phenomena with which we are concerned too often are inferred states that derive as much from the viewpoint of teachers and psychologists as they do from the pupils themselves. Moreover, the language used to denote and define the dimensions of such things as subject achievement, intelligence, study habits, and citizenship usually is of the vernacular. The words then have broad and varied connotations. For example, these might be the dimensions of achievement as defined in a general science course of study: factual knowledge, critical-mindedness, habits of accuracy, understanding of relationships, and scientific attitudes. It is exceedingly difficult to know exactly which of the many possible meanings for any of these words is intended, let alone to differentiate one dimension clearly from any other.

So, in arranging educational phenomena for measurement, an essential first step often must be to define their dimensions in unequivocal terms. As we shall see, this generally means to define them in terms of simple and observable actions and the observable attributes of the actions. For example, "critical-mindedness" might need to be restated as "points out errors in evidence or in arguments from evidence." We can listen to or read the errors a pupil detects and we can count *how many* and identify *what kind*.

4. *If a group of phenomena are to be measured with respect to a dimension, they must differ with respect to the dimension.* Variation is inherent in the act of measurement. In fact, dimensions often are called *variables*. We measure the intelligence of pupils because they have *different degrees* of intelligence. We measure their achievement in social studies because they exhibit different degrees of achievement.

If all members of a group are alike with respect to some property, its measurement has no significance. For instance, all pupils in attendance at school are alive. We are aware of no differences among them as to their degree of sentience; so teachers do not customarily measure *how alive* they are. Physicians and undertakers are, of course, concerned with tests for "livingness" because they inevitably see those who are dead.

We are most prone in measurement to think of the differences among

phenomena as lying along some sort of a continuum or of constituting a continuous variable. Intelligence, height, weight, subject achievement, motor skill, are examples of dimensions in which variation is thus a matter of degree. There are, however, other types of variation. We have dimensions in which variation is discontinuous. Enumeration (how many) is one of these and it pertains to anything that has parts or distinct units. For other dimensions the variation is not serial or a matter of degree; it is simply one of type or category. For example, there are dichotomous dimensions such as sex and many-classed ones such as race, nationality, and political affiliation. The variation in some categorical dimensions has to do with place in a hierarchy or other systematic arrays; for example, the "specie" of a life form and the "somatotype" of a person's physique. While such instruments as rulers and meters, which yield measurements along a continuous scale, are not applicable to these other types of variation, procedures that record the differences in appropriate ways are applicable, and the dimensions are measured only as the differences are detected.

The fourth essential condition of what we can measure is, then, variability. Dimensions exhibiting the more easily discernible and the more consistent differences are, of course, the more precisely measurable.

5. *Measurable dimensions must produce highly similar reactions among many unrelated and impartial observers.* Finally, have you ever tried to measure a ghost? Some persons say they have seen and heard them. They are defined rather clearly by these persons and they are said to differ with respect to size and temperament. The British Society for Psychical Research has for some time been attempting to measure ghosts, but the results of the measurements are not yet satisfactory even to many members of the Society.

On the other hand, have you ever tried to measure a motion-picture ghost cast on a screen by a projector? If you were to try, your appraisal as to height, luminosity, etc., probably would be accepted, as long as you used a rule, a light meter, etc.

Now why is it that projected images of movie ghosts can be measured and ghosts themselves cannot be? They are both observed by some persons. They are both rather clearly defined, and they both exhibit variation. You say a ghost isn't real, that ghosts are just superstitions? We agree. But suppose 99.9 per cent of all the people are blind and only 0.1 per cent can see. If these 0.1 per cent see projected images and talk about them to the blind who, being an overwhelming majority, publish all the scientific journals, might not projected images also be thought a superstition?

So, what essential characteristic does distinguish the ghost from the projected image as to measurability? It is thought to be simply that many unrelated and impartial observers *agree* essentially as to what they observe when they observe a projected image; and, of great importance, *all*, or nearly all, persons with 20/20 vision do see something when an image is projected. With ghosts, however, only related observers agree even a little, unrelated ones dis-

agree a great deal, and impartial observers generally never see *anything* except where the ghost is found to be a projected image or an optical illusion.

This final condition of measurability for a dimension—impartial observers must agree in their reactions—has great importance for educational measurement. The appraisal of certain things is unreliable, even impossible, just because teachers agree neither as to their dimensions nor as to the status of pupils relative to these dimensions. Character, personality, citizenship, appreciation of art are among the phenomena that often fail to produce agreement among observers. Consequently, before one attempts to measure them, he must find or devise dimensions for which there can be some consensus of reaction.

In the preceding paragraphs we have developed five conditions of measurability. To recapitulate, measurement symbols can be validly assigned to phenomena in the degree to which their dimensions:

1. Are possessed in common by a group of phenomena
2. Provide sensory data
3. Are clearly defined
4. Manifest variation
5. Yield equivalent reactions from unrelated and impartial observers

In applying the second, third, and fifth of these criteria to educational phenomena, it is necessary to consider that they represent poles or extremes of continuums, at which other extremes are conditions that prevent measurement. In the middle of these continuums lie conditions intermediate between accurate measurement and no measurement. For items found here, it will be necessary to decide whether they are sufficiently observable, defined, and productive of consensus in observers to be considered measurable dimensions. These considerations are made graphic in the following chart.

<i>Obviously Measurable Dimensions</i>	<i>Measurable Dimensions?</i>	<i>Obviously not Measurable Dimensions</i>
Differentiated sense data		No or vague sense data
Well defined		Not defined
All unrelated and impartial observers agree		No unrelated and impartial observers agree

In preparing dimensions of educational phenomena for measurement, it may be assumed that they will meet the conditions of being applicable to a group and of manifesting variation. On the other hand, it should be assumed that they will only approximate the other three conditions and that many will fall so short of the conditions as to be unmeasurable in their given form. For this reason, the tests of sensory data, definition, and consensus among observers must be rigorously applied to educational phenomena before their

measurement is attempted. *In every case* dimensions should be so construed that the maximum of extensive and appropriate sensory data is provided, that their definitions are as clear as possible, and that maximum agreement may be expected among observers. If in your view any dimension should still fall too short of one of the conditions for accurate measurement, its measurement should not be attempted; *for once a measurement symbol is applied to an individual, there is a presumption that something has been measured.* And we have attempted to show that *nothing* may have been measured unless the five conditions of measurability have been approximated.

Obviously, these conditions are interrelated. Where one is met there is likelihood that the others are as well, and efforts to correct one may correct others. For example, *clear definition* is possible only with some degree of *sensory data*. It is possible, then, to indicate one general method of defining phenomena and their dimensions so the three critical conditions of sensory data, definition, and consensus among observers may be approximated. This is:

1. To define the phenomenon in terms of its dimensions
2. To define the dimensions in terms of specific behaviors and their observable attributes

This method is explained and stressed throughout the book in connection with discussions of measuring procedures and the many focuses and uses of measurement in education. Here we will include only a simple application of the process to illustrate its effectiveness.

"Literary appreciation" is one of the many aspects of pupil achievement whose evaluation continues to exasperate teachers. One reason for this situation may be that "literary appreciation" often is not clearly defined and has no clearly indicated dimensions. To measure it requires first that it be defined in terms of its dimensions or properties. One such definition might be: Literary appreciation means the degree to which a pupil:

1. Understands the structure and symbolism of great writing
2. Is aware of the social significance of poetry, prose, and drama
3. Discriminates between popular and classic works ⁵

The basic dimensions of literary achievement, according to this definition, are the elements listed. But, as stated, none of them is clearly defined, they offer no sensory data, and observers would argue about them. To redefine them so as to correct these shortcomings, we must state them in terms of actions in which a pupil manifests his status re the dimensions. For the third dimension, "discriminates between popular and classic works," such behavioral redefinition might be in part: "*Choice* of titles for free reading; *recall* of titles, characters, and situations in conversation and writing; *comparisons* of current writings

⁵ There are many other possible definitions of literary appreciation. An excellent one is published in the *Thirty-Seventh Yearbook of the NSSE*, Part I, Bloomington, Illinois, Public School Publishing Co., 1938, pp. 114-115. See also Chapter 10, page 254 here.

with those deemed classic." This redefinition seems to be reasonably clear. It is possible to listen to or to read what titles pupils recollect and what comparisons they make. Moreover, observers are likely to agree about the choices, recollections, and comparisons. Finally, pupils' status in respect to these operational dimensions is a function of *how many* classic versus nonclassic titles were read, *how many* classic versus nonclassic titles, characters, and situations are recalled, and *how many and what kind* of comparisons do they make between current and classic writings. Thus, in counting *how many* and in stating *what kind*, we indicate the degree to which the pupils manifest this given aspect of literary appreciation.

Inferred Dimensions and Constructs

We have discussed the fact that measurement may be directed only to observable dimensions. Yet, as we have noted, many of the things that educators need most to appraise are abstractions about behavior or are covert, unobservable states of mind and feeling. Often to these so-called "intangibles" are ascribed properties that are themselves unobservable. Scientifically speaking, any such properties constitute inferred dimensions. Since inferred dimensions provide no sensory data, they cannot be measured directly. They can, however, be measured indirectly by finding or devising *observable* dimensions that are related to the inferred ones.

Intelligence is one of the many educational phenomena that have inferred dimensions. It is customary to speak of inductive reasoning as an aspect (or dimension) of intelligence and most intelligence tests attempt to measure a child's ability to reason inductively. Yet "inductive reasoning" as such is not observable. We haven't as yet been able to get inside the mind and directly measure what goes on there when a person reasons. Hence, the dimension is an inferred one. We measure it by observing a child at tasks that we say *require inductive reasoning*: tasks such as making up rules to explain what has happened, solving codes, working out ratios, etc. Sometimes an observed dimension is measured so often to determine the same inferred one that we become interested in it for itself. We treat it as if it really belonged to the phenomenon. Those who give individual mental tests talk about "digit span." This is how many numbers a child or an adult can recall, having heard them just once. By continued usage, "digit span" has become nearly as much a dimension of intelligence as the "memory" to which it relates.

Inferred dimensions may be measured as accurately as any if they are properly inferred and if the related dimensions to be directly measured are properly chosen. Because of this we are able to evaluate pupils' knowledge, their personality, their intelligence, and their attitudes, all of which are themselves unobservable. In Chapter 6 and in Chapters 11 and 12 we shall see that the construction of achievement tests is largely a matter of keying the observable responses of pupils to test questions on the one hand to the inferred dimensions of their knowledge of school subjects on the other.

Inferred dimensions often belong to a type of phenomena called *constructs*. Intelligence and personality, just mentioned, both are constructs. So are the atom, radio waves, the subconscious mind, and many other things. A construct is an explanation, not a thing. It is not a rock; it is a geologist's theory of the molecular structure of that rock.

Very simply, constructs are symbolic maps where words and their interrelationship, numbers and their interrelationship, represent the structure or process of unobservable biological and physical states. They are based on and serve to explain the observable data for which we assume some unseen or underlying causation. Like maps, constructs summarize; they boil things down to a small scale. Like maps, constructs leave out details, they abstract, omit, and select. Neither maps nor constructs need to *look like* the terrain or structure they explain; they use conventional symbols.

Finally, both maps and constructs facilitate predictions about observable things. Open a highway map of Missouri and see how far it is from Kansas City to St. Louis: 257 miles. At 50 miles per hour, with allowance for stops and traffic, that should take $6\frac{1}{2}$ hours. So tomorrow you allow $6\frac{1}{2}$ hours to go to St. Louis from Kansas City and you get there in $6\frac{1}{2}$ hours. With some constructs, electronics for example, one can make fantastically accurate predictions. Vacuum tubes are built to make a certain pattern on the screen of a specific television set and they do just that, all on the basis of an engineer's specifications. Yet the engineer need never see the tube, nor the set, but only some electrical and mathematical symbols on paper. With some constructs of educational significance, there is much less accuracy in predictions. Knowing that ten children aged six have IQ's of 90 and ten other six-year-olds have IQ's of 110, we would predict that the latter group, as a group, will learn to read faster and read more in the primary grades than the former group. We order books and plan instruction accordingly, and our prediction is fulfilled. However, for just *one* child of low IQ and *one* child of higher, our prediction would be less confident and our mistakes more frequent. Unfortunately, our current constructs of intelligence are sometimes like the maps of the "New World" in the time of Amerigo Vespucci. According to legend, with his maps as navigational guides, it was not at all difficult to miss a continent, let alone an island.

The inferred dimensions of a construct are the properties which the construct needs to have to explain the observable data to which it relates. The test of their validity is the accuracy of the predictions they support. For example, "memory" is a valid dimension of intelligence to the extent that accurate predictions may be made about the academic success of pupils as the result of measurements of their ability to remember.

Even though an inferred dimension is a valid one, it still needs to be measured accurately and this first entails selection of appropriate observable dimensions. With respect to "intelligence" and two of its dimensions, memory and reasoning, history has shown us what are and what are not appropriate

observable dimensions. In the eighteenth century, and even in the nineteenth century, it was not uncommon to assess memory and reasoning by measuring the contours of the skull, by observing how high was the forehead or how wide-set the eyes, sometimes even by asking the date of birth and the concurrent status of the zodiac. Now, to test for memory and reasoning, we ask children to recall numbers, words, and pictures and to do problems in arithmetic, work puzzles, run mazes, and the like.

Principles in the Selection of Dimensions

For the most part, the phenomenon to be evaluated will indicate the dimensions to be measured. If it is spelling ability, *what* and *how many* words can the pupil spell? If it is a question of facility at wiring in a shop situation, the dimensions are *how long* the student requires to complete the job, the *errors* in his connections, the *quality* of his soldering, etc. Sometimes, however, the phenomenon itself does not immediately disclose the dimensions appropriate for measurement. In such cases it is well to have in mind certain principles that are relevant to their selection.

In the first place, dimensions should be selected that relate directly to the purpose of the measurement. For example, the sponsor of the school paper, who wishes to pick the student staff, might wish to measure the *variety* of adjectives and adverbs that prospective reporters can use. On the other hand, the secretarial teacher, in choosing students to do stenographic work in the school, might not be interested in such a factor as this but perhaps only in the frequency of grammatical and spelling *errors*. Yet both teachers would be measuring the same basic thing, composition.

Secondly, dimensions should be selected in terms of the precision required. Any measurement purpose entails a minimum degree of precision in results if the purpose is to be accomplished. Precise appraisal is partially a function of the dimensions measured and consequently we need to select the dimensions that give the required precision. To illustrate this point, if the third-grade teacher needs only to distinguish three levels of reading ability relative to the Readers she uses, she then may be concerned only with rate and comprehension for this material. The reading diagnostician, however, needs to know the exact status of each pupil for all types of reading; so he appraises the dimensions of rate and comprehension for each of several types of material, and additional dimensions (*recognition techniques, sight vocabulary, etc.*) as well.

It sometimes occurs that the measurement task may not be accomplished satisfactorily through the measurement of *any* dimensions currently known for the phenomenon. It is proper and essential, in this event, to find or invent new dimensions through imagination, experimentation, and theorizing. New dimensions are commonplace for the physical scientist. In fact, science has advanced as scientists have had the temerity to name and the tenacity to measure new properties. Think of Curie and "radiation," Newton and "gravity," and some unknown ancient sea captain and the "polarity" of the earth.

In education, the finding and inventing of new dimensions is less remarkable, but a scientific approach to education is a very young idea. However, all these very useful dimensions of pupil behavior are new: intelligence, needs and drives, social status, group cohesiveness, ego defenses, aggressiveness, compensation mechanisms.

Finally, dimensions should be selected in light of the standards to be used in evaluation. We observed in the overview to this section that effective evaluation requires the comparison of a phenomenon—achievement in English, knowledge of history, swimming ability, anything—with a standard. The standard normally consists of one or more levels or classifications of status to which are assigned one or more symbols of value or worth. For example, an evaluative standard in typing might be: “Less than 20 words per minute, unsatisfactory; 20 to 40 words per minute, fair; 40 to 60 words per minute, good; 60 and more, professional.” In Chapter 9, the development and use of evaluative standards are discussed in detail, and in Section II attention is given to the standards employed in various school subjects.

Obviously, a standard must be expressed in terms of one or more dimensions of the phenomenon to which it refers. It follows then that measurement that is to lead to evaluation *must* relate to precisely the same dimensions as are cited in the evaluative standard. Any less will make the evaluation indeterminate or impossible; any more will be superfluous. In the typing example, you could use the standard only if you gave a speed test or otherwise measured the dimension of speed; and there would be no point to measuring accuracy if it were omitted from the standard. Obviously, if the standard referred both to speed and accuracy, both of them would need to be measured.

Some Dimensions Common to Most Educational Phenomena

In later chapters we will cite many of the varied dimensions of specific educational phenomena. Here it may be useful to mention several general dimensions likely to be common to most of them. Knowledge of these should help in understanding the particular dimensions given for some specific subjects and in determining the measurable dimensions of others.

Since all educational phenomena have parts or aspects, the *identity*, *number*, and *organization of components* are common and important dimensions. In measuring a student's ability in composition, for example, the types of sentences, clauses, and phrases he uses are to be identified, the number of varied constructions is to be determined, and the ways in which paragraphs are organized is to be noted. “Tests” in history and geography invariably seek to probe what facts (*identity of components*) and how many facts (*number of components*) a pupil knows and how he has organized these facts (*pattern among components*).

A fourth common dimension of educational phenomena is *time*. How long this behavior has lasted is a frequent concern of the teacher. Speech teachers give attention to this dimension as they listen critically to a student's

talk and tell him to sustain the sound of given words longer or to speak more crisply.

Rate or *speed*, a fifth common dimension, is well known to reading and business instructors and is formed of two other dimensions, time per given component. Forty words per minute in typing, fast tempo in music, 300 words per minute in reading, are examples of measurements of *rate*. A sixth common dimension, *frequency*, is an enumerative counterpart of *rate*. Examples of the dimension are: number of errors per 100 words, number of pupils who are twelve years old, and number of times Bill or Mary was disorderly this week.

Finally, *error* itself is a dimension common to pupils' achievement in all school subjects. Certain ways of spelling words, certain answers to multiplication problems, certain name-date associations, and a myriad of other things are called *errors* because they are all departures from how it is or how it should be. Since rectitude of performance is a primary goal of education, and since pupils have never ceased to depart therefrom, *error* must be designated as a general dimension of achievement.

The choice of dimensions of pupil achievement in those school subjects that we call areas of knowledge or understanding poses a particular problem. While there is a general agreement that science, history, geography, etc., should be taught, there is less agreement on what should be taught in their name. And there is practically no agreement on *what* should be *measured* when a student's proficiency in any knowledge subject is to be judged. Much of this difficulty may stem from the fact that the "knowledge" must be an attribute of the pupil, if we measure him relative to it. On the other hand, that which we think of when we try to measure "knowledge" often is apt to be the content of the books and other documents comprising the "knowledge" we try to teach him. The books contain words and other symbols arranged in logical and static ways, while the pupil contains action and feeling tendencies arranged psychologically and dynamically, if he may be said to "contain" anything at all.

However the problem arises; its solution requires that any subject of knowledge or understanding be clearly analyzed from the viewpoint of measurable dimensions before any measurement is undertaken. The general dimensions we have just discussed offer a starting point in this analysis. The possible components of the pupils' knowledge, their identity, number, and organization may be construed as the expressed facts and concepts, laws, etc., which the subject encompasses and the logical or chronological relationship among them. The dimension of error could refer to errors in the pupils' grasp of such facts, concepts, laws, etc., as against their representation in books and other reputable documents.

In addition, though, if understanding of something is to be measured adequately, dimensions should be selected in terms of the psychology of the pupil, not just in terms of the logic of the subject. The pupil learns the elements of the subject in various ways: he relates them to other experiences; he applies

them to new situations, modifying them as he does so. His knowledge at any moment is a function of his mind in action. It is not simply a reflection of some portions of books and lectures with an occasional distortion.

Consequently, pupil knowledge or understanding of any subject necessarily has such dimensions as interpretation, application, and synthesis of facts, as well as that of facts recalled. It has dimensions that have to do with the hierarchical arrangement of ideas as well as their number.

Later in the text a number of school subjects are examined in the light of this viewpoint toward the dimensions of knowledge and understanding. A recent remarkable publication, *Taxonomy of Educational Objectives* (2) presents a general classification of educational goals tantamount to a statement of the possible basic dimensions of student achievement. The various classes of goals (or dimensions) stated for knowledge and understanding are given in terms of pupil thought and action and not just the logic of a "subject." Furthermore, the authors describe appropriate methods of testing to determine the status of students in respect to each of the goals. While it may have some flaws, this taxonomy could be an extremely useful basis for devising the measurable dimensions of achievement in particular subjects. The categories the taxonomy presents for the "cognitive domain" are as follows:

Knowledge

- 1.0 Knowledge
 - 1.10 " of specifics
 - 1.12 " of specific facts
 - 1.20 " of ways and means of dealing with specifics
 - 1.21 " of conventions
 - 1.22 " of trends and sequences
 - 1.23 " of classifications and categories
 - 1.24 " of criteria
 - 1.25 " of methodology
 - 1.30 " of the universals and abstractions in a field
 - 1.31 " of principles and generalizations
 - 1.32 " of theories and structures

Intellectual Abilities and Skills

- 2.0 Comprehension
 - 2.10 Translation
 - 2.20 Interpretation
 - 2.30 Extrapolation
- 3.0 Application
- 4.0 Analysis
 - 4.10 Analysis of elements
 - 4.20 Analysis of relationships
 - 4.30 Analysis of organizational principles
- 5.00 Synthesis
 - 5.10 Production of a unique communication
 - 5.20 Production of a plan, or proposed set of operations

- 5.30 Derivation of a set of abstract relations
- 6.0 Evaluation
- 6.10 Judgments in terms of internal evidence
- 6.20 Judgments in terms of external criteria

Summary

Before educational phenomena may be measured, they must be construed so as to be susceptible to measurement. Often, as currently stated, they serve to confound rather than to facilitate measurement. Phenomena may be measured only in terms of their dimensions, properties, or attributes. Moreover, several basic conditions must be approximated if any dimension is itself to be measured. The dimension should:

1. Be common to a group of phenomena
2. Provide sensory data
3. Be clearly defined
4. Manifest variation and
5. Produce highly similar reactions among many unrelated and impartial observers

Dimensions may be inferred as well as directly observed. Constructs (i.e., personality, intelligence, knowledge) are composed of such inferred dimensions. To measure them, each inferred dimension must necessarily be related to an observable one. Dimensions should be selected for measurement according to the nature of the phenomenon, the purpose of the measurement, the degree of precision desired, and the standard to be used in evaluation.

Certain general dimensions are likely to be common to most phenomena of educational interest: identity, number, and organization of components, time, rate, frequency, and error. For the phenomenon of "knowledge" or "understanding," it is essential to select dimensions that relate to the mental activity of pupils as well as to the logic of the subjects they are to know.

EXERCISES

1. Define the following terms as they are used in this text:

Measure
Evaluate
Phenomenon
Dimension
Sensory data
Inferred dimension
Construct
Standard

2. Assess the relative measurability of these basic dimensions of a pupil in terms of the five conditions of measurability:

Age
Height
Weight
Knowledge
Intelligence
Interests
Background
Personality

3. Indicate several measurable dimensions for (a) skill in a given activity, (b) knowledge of a given subject, and (c) attitude toward school.

4. What would you say are the measurable dimensions of the ability to think effectively?

BIBLIOGRAPHY

1. Anastasi, Anne, "Concept of Validity in the Interpretation of Test Scores," *Educational and Psychological Measurement*, 10:67-78, 1950.
2. Bloom, B. S. (ed.), *Taxonomy of Educational Objectives*. New York: Longmans, Green and Co., 1956.
3. California State Department of Education, "Evaluating Pupil Progress," Bulletin No. 6, 1952, 184 pp.
4. Cattell, R. B., "Objective Measurement of Dynamic Traits," *Educational and Psychological Measurement*, 10:224-248, 1950.
5. Comrey, A. L., "Mental Testing and the Logic of Measurement," *Educational and Psychological Measurement*, 11:323-334, 1951.
6. Deese, James, *The Psychology of Learning*. New York: McGraw-Hill Book Co., Inc., 1952, chap. XIV.
7. Greene, Edward B., *Measurements of Human Behavior*. New York: Odyssey Press, 1941, chaps. I and II.
8. Lindquist, E. F. (ed.), *Educational Measurement*. Washington, D.C.: ACE, 1951, chaps. I, II, III, VI, and XIV.
9. Loevinger, J., *Systematic Approach to the Construction and Evaluation of Tests of Ability*. American Psychological Association, 1947.
10. Michaelis, J. U., "Current Practices in Evaluation in City School Systems," *Educational and Psychological Measurement*, 9:15-22, 1949.
11. National Society for the Study of Education, *The Measurement of Understanding*, Part I, 45th Yearbook, 1946, chaps. III and IV.
12. Scates, D. E., "Some Problems Connected with Evaluation," *Journal of Educational Research*, 45:599-608, April, 1952.
13. Suchman, E. A., "Logic of Scale Construction," *Educational and Psychological Measurement*, 10:79-93, 1950.
14. Thomas, R. Murray, *Judging Student Progress*. New York: Longmans, Green and Co., 1954, chap. II.
15. Travers, R. M. W., "Rational Hypotheses in the Construction of Tests," *Educational and Psychological Measurement*, 11:128-137, 1951.
16. Wrightstone, J. W., "New Emphasis and Broader Scope for Tests and Measurements," *Nations' Schools*, 43:39-41, March, 1949.

CHAPTER 3

PROCEDURES OF MEASUREMENT IN GENERAL

We now have discussed two major aspects of the measurement of educational phenomena. In the first chapter we examined the forms of symbolic expression that measurements may produce: classification-description, rank or order, and scaling. We determined in the second chapter that symbols (numbers or, if not, words and letters with precise definition) appropriate to any of these forms may be assigned to any phenomenon you wish to measure—but only as long as the phenomenon *is* measurable. To help gauge the measurability of phenomena, five conditions of measurability were established. If it is to be measured, a thing must have one or more dimensions (property, quality, aspect, etc.) in common with other things. These dimensions must provide us with some sensory data, be discrete and well defined, manifest variation, and produce highly similar reactions among unrelated and impartial observers. It was noted that the measurement of physical phenomena rarely involve the application of these conditions, so well do the familiar physical dimensions meet the conditions. On the other hand, behavioral phenomena, the things of primary interest to teachers, must invariably be examined for measurability before their measurement is attempted.

At this time, we shall give attention to the third and most time-consuming aspect of measurement: the procedures by which measurement is accomplished. If we were concerned with measuring speed, height, weight, and other physical dimensions, we would start to talk about speedometers, yardsticks, balances, photometers, calipers, spectrometers, etc., and the techniques for the use of any of them. Since, though, we are to deal exclusively with behavioral phenomena, our attention will be on the devices we use to appraise behavior: tests, rating scales, observation schedules, and the like. In the four succeeding chapters we will present the nature and significance of the following procedures together with principles for their efficient use: observation, product analysis, free-response procedures, and guided response techniques. Before we turn to them, however, it is necessary that we understand something about behavioral measuring procedures in general.

Function of Measuring Procedures

When, as teachers, we give arithmetic or science tests to our pupils we have to score the papers. We count the number of answers wrong or right and place

a corresponding number or letter on the papers. When, in an English class, we ask the pupils to write a short composition so that we may see how good is their usage, we similarly mark their submissions and place an appropriate number or letter on the compositions. Now, this number or letter or word *is the measurement*. The purpose of the tests and of the writing of the compositions has been only to obtain *these numbers or letters*. Thus, the burden of a measuring procedure is the assignment of proper measurement symbols. These characterize the status of the pupils in regard to arithmetic, science, or composition, which was our purpose in measurement.

Obviously then, the things we may be accustomed to think of as constituting measurement—true-false tests, blue books, rating forms, and all the other paraphernalia of educational measurement—are only the means of measurement. If the proper measurement symbol may be assigned *without using the procedure*, the procedure is of no use. If a student's achievement in history can be symbolized as validly, reliably, and efficiently by just listening to him recite as by testing him, there is no point to testing him. On the other hand, we shall see presently that the assignment of the proper measurement symbols usually necessitates some sort of set procedure. Hence, our tests, our handwriting scales, our interest inventories, and our aptitude batteries are indispensable to measurement if they are not synonymous with it.

Basic Properties of Procedures

The fact that the phenomena of primary interest in education are behaviors or psychological states poses problems for measurement procedures. It is no trick at all to assign the proper scale number to a boy's height. Stand him alongside a rod or wall lined off into feet and inches and read the point with which the top of his head intersects. Five feet, one inch, and that's all there is to it! But to measure his intelligence, his knowledge of arithmetic or his attitudes toward school may require more complicated devices and a great deal more ingenuity. Behavioral or psychological entities are usually processes, many of their components are unseeable, they do not stand still or have any mass, they change very rapidly, and they amount to more than can be measured at any one time.

To accommodate to these conditions and yet to produce reasonably accurate measurements, the procedures of behavioral measurement employ *standard stimulations*; they attempt to elicit *standard differential responses*; they use *standard analysis systems*; and they engage in *sampling*. Sampling and standard analysis are inherent in all types of procedure but the other devices may or may not be the property of a given procedure.

STANDARD STIMULATIONS

The obvious characteristic of what you know as a test is that it provides the same stimulation to all pupils. Both the true-false question and the essay question are intended to provide the pupils who read them with the *same*

sensation so that their *different* responses may be compared. Obviously, unless the pupils respond to the same thing, the significance of any comparison made among them is unknown. And, as we know, comparison among pupils is necessary in many instances of behavioral measurement.

What standard stimulations are presented is, of course, determined by what responses are desired and these in turn by what dimensions are to be measured. For example, the statement, "Columbus discovered an island in the Caribbean Sea in 1492," probably is used to elicit a "Yes" or a "No" from pupils and the "Yes" or "No" is desired because it is thought to be keyed to the *facts* known by a pupil about American History. The purpose of the question, "What is the relationship between on-shore winds, coastal mountains, and rainfall?" is to obtain a sentence, a paragraph, or more from students that should reflect their different understandings of the interrelationship among climatic factors.

The types of standard stimulations employed in educational measurement are many. They will be presented in connection with free response techniques and, in particular, with guided response techniques. Whatever procedures are employed, behavioral measurement requires the occurrence of behavior or evidence of a behavior, and standard stimulations are devices to obtain comparable behaviors or evidences thereof from a number of pupils.

STANDARD DIFFERENTIAL RESPONSES

If the most obvious characteristic of a test is its inclusion of standard stimulations, a second, but equally important characteristic, is its attempt to elicit standard yet differential responses. The necessity of differential responses occurring is axiomatic. We established earlier (Chapter 2, page 22) that a measurable dimension is one in which variation is manifest. Consequently, any procedure for measuring a dimension must be sensitive to or productive of differences. The least differential responses are obtained by the very familiar true-false items; the most, by such guided response items as essay questions and tell-me-a-story pictures. The true-false variation is only twofold, the least possible, while, within limits, no two responses to an essay question are ever identical and, hence, the variation of responses to them is as great as the number of students tested.

By "standard" is meant simply that the same words or actions given by different pupils are to mean the same thing and that the options of response possible in a given situation are known and have predetermined significance. Multiple-choice questions exemplify test items that elicit highly standard responses. Consider the following item:

- A. The achievement of the Greeks was relatively greatest in:
 1. war
 2. social reform
 3. government

4. laborsaving machinery
5. navigation charts

A pupil may answer this only by placing a number, 1 to 5, in the space provided, and thus the nature and number of possible responses are strictly controlled. One of the numbers means that the pupil has correct knowledge concerning this aspect of Greek civilization. All others mean that he has incorrect knowledge or no knowledge (guessing). Hence, the options have predetermined significance and they are assumed to have the same meaning no matter which pupil uses them.

Essay questions, on the other hand, produce responses much less standardized. The words used by different pupils are assumed to mean the same thing but the significance of different types of response is only roughly predetermined and the number and nature of possible responses is unknown. As we shall see presently, this relative lack of standardization tends to make essay questions and other free response items more difficult to score and inherently less reliable than guided response items.

Obviously, the more standardized are the answers to test questions, the more comparable are the answers given and the more impersonal their scoring. True-false, multiple choice, matching questions, and other guided response items yield test scores that may be ranked and can be marked mechanically or electrically, the ultimate in objectivity. The answers to essay questions given by several pupils are not as amenable to ranking and they are not susceptible, to machine scoring. Even with a well-devised system for marking, the examiner must make many judgments as to the meaning and significance of phrases.

Standardization of responses is gained, though at the expense of differentiation. For attributes wherein pupils are known to manifest great variation (intelligence, personality, musical ability, etc.) objective tests with their rigidly standardized responses have limited usefulness.

STANDARD ANALYSIS SYSTEMS

The *sine qua non* of procedures of educational measurement or any measurement is a standard analysis system. Some are very simple, the "key" to a science test or an observation sheet with captions and space for writing. Others are very complicated, a factor-count schedule for a projective personality test or the scoring standards for an individual intelligence test. But, simple or complex, the function of a standard analysis system is to insure that the same measurement symbols are assigned to phenomena of similar status; that different symbols are assigned to phenomena of different status; and that all phenomena purportedly measured in regard to a given dimension are each measured in the same way for just that dimension.

To this end varieties of techniques are employed for detecting, recording, classifying, counting, and comparing the proper dimensions of the behavioral phenomenon being measured. An observation check list helps the teacher look

for and, thus, detect in a systematic way the pupil actions he considers relevant to citizenship. The record forms and rating blanks used in individual intelligence and personality testing are simply the means by which the examiner records his observations of all those he tests in a standard fashion. A test key is a good example of a classifying device. It permits the automatic assignment of items to categories of right and wrong, or whatever other categories are desired. An electrical scoring machine not only classifies responses automatically, but also counts the number of responses in each category. Perhaps the most familiar illustration of a comparison device is a Handwriting Scale. In this, as you doubtless recall, graded samples of penmanship are provided for comparison with a specimen written by a pupil. Different samples are tried until the one is found that is most like the pupil's specimen. The specimen is then assigned the rating of that sample.

Additional devices frequently are used to convert verbal data into numerical data, to convert raw or first numbers into those having class, rank, or scale significance, and to derive measurements of a group from the measurements of the individuals within the group. The latter two conversions are the burden of the statistical procedures described in Chapter 7. The tables and profile sheets of standardized achievement tests (see page 125) are the means by which a teacher converts the number of questions a pupil answered correctly (raw score) into a grade index, a percentile equivalent, or sometimes a scale number.

SAMPLING

The last general feature of measuring procedures we need to consider is that of *sampling*. The idea of sampling is simple. You measure a small portion of something and then make statements about all of it. But the implications of sampling for measurement are many, varied, and not always obvious. Therefore, to explain the concept properly, we shall use a very tangible illustration.

Suppose that a housewife, who has just bought an apple pie from a small bakery, eats a piece and then says to her husband and children, "This is a good pie, have some." The woman has made a declaration about all the pie, but she has tasted only a piece of it. Since a pie made by a bakery tends to be uniform throughout, it is highly probable that her husband and children will find it good also. But it is *not certain* that they will do so. The side she tasted may contain more or less of the fruit than the other side. Her piece may have received more or less heat, and so on.

Now suppose the housewife does find all the apple pie to be good, and because of this she considers the idea of buying all the pies the bakery has produced that day (which happens to be Tuesday). Will that be wise? Perhaps the bakery's berry pies contain overripe fruit, perhaps in their meringue pies powdered eggs have been used.

Finally, suppose that the housewife does buy the bakery's entire output of pies for Tuesday, stores them in her freezer, and eventually she and her

family eat them all and find them uniformly delicious. Should she then decide to buy all this bakery's pies each Tuesday? She should only if she can be sure that the oven's temperature next Tuesday will be at exactly the same temperature as this Tuesday, that the dough will be mixed just as long, that the fruit, custard, and chocolate will be just the same in quality, and so on. Obviously, she cannot be sure.

In this illustration we may observe the several facets of sampling and from our observation we may derive some basic principles that affect all measurement procedures. Figure 1 shows them in graphic form.

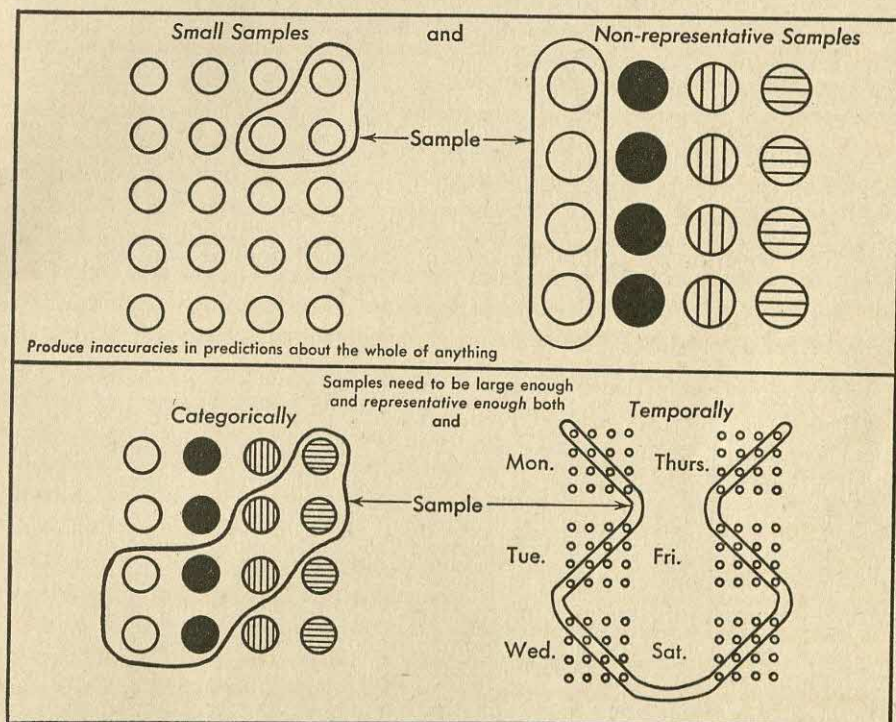


Figure 1. Principles of sampling.

1. *Appraisals that involve sampling are estimates or predictions only. Until the housewife has eaten all the pie, her statements about it are subject to error.*

2. *Estimations based on sampling are least accurate when the sample is a small proportion of the whole (e.g., one sixth of the pie) and when the sample is not representative (e.g., only apple pie eaten, but berry, custard, etc., to be purchased). Conversely, estimations based on proportionately large samples (five pieces of the pie) and on representative samples (all varieties tasted) are most accurate.*

3. *Sampling may be categorical (a fraction of a pie, one or several pies*

out of many). *It also may be temporal* (Tuesday's pies as a sample of pies baked during the year). Both kinds of sampling usually are involved in educational measurement.

Educational measurement entails sampling for several reasons. First, the phenomenon to be measured may consist of an unlimited number of factors. A pupil's intelligence or the achievement of an eleventh-grade class in physics are examples of such phenomena. Second, to measure all the cases in a given population¹ may be too costly or time-consuming. For example, to gauge the home conditions of each pupil in a large city school district would require a large staff of sociometrists and thousands of dollars. Finally, the act of measurement may consume the items that are being measured and, hence, only a sample may be used. The chemist with his flame and precipitation tests for unknown compounds exemplifies this reason for sampling.

The degree of indeterminacy inherent in measurements because of sampling may be stated mathematically. The relatively simple mathematical procedures available for determining such "sampling errors" are explained in Chapter 8, pages 169–172. The specific applications of sampling to the several types of measuring procedure are presented later in context with the procedures.

Criteria for Procedures

Just as we found that a dimension to be measured must first be gauged for its measurability, so we should recognize now that a procedure of measurement should not be used unless it is judged adequate for the purpose. Obviously, everything called a test does not test with equal accuracy. In the experience of each of us have been instances when we felt that a test was unfair, too short, neglected certain aspects of the course, etc. Equally distasteful in our recollections are the occasions when we ourselves judged people wrongly or sized up situations improperly on the basis of faulty observations.

As a basis for discriminating between relatively good and relatively bad procedures, we may use three criteria: validity, reliability, and efficiency. These, furthermore, are the criteria that should govern the construction of any test or the devising of any other procedure of measurement.

VALIDITY

According to the customary viewpoint, the capability of a test or other measuring procedure to *measure* what it *purports to measure* is called validity.²

¹ Population is used here to mean any aggregate of items—not just a group of people.

² Another approach to validity has been expressed by Cronbach (4:48), "A test is valid to the degree that *we know* what it measures or predicts." From this point of view, the measurement focus of the test is indeterminate until we inspect its items or until we see with what other measures its results correlate. The title or announced objective of the test is of less importance. From this point of view, however, the practical consequences for the ordinaries of school measurement seem to be about the same as they are for the first and more usual definition, i.e., concern over factorial purity, clear definition of dimensions, elimination of irrational response determiners, etc.

The extent to which it does this is its *degree of validity* and when such degree is expressed by a number it is called a *coefficient of validity*.³ For example, if a test of "mathematical achievement" differentiates among pupils on the basis of their differing skill and knowledge in mathematics and that only the test is a valid test. If another test of mathematics measures mathematical skill plus something else, intelligence, let us say, the latter test has a *smaller degree* of validity than the first.

Validity in a measuring procedure is a function of several things.

1. Before any consideration is given to the procedure itself, validity involves *a clear definition of the phenomenon and its dimensions for which measurement is to be sought*. This description should so state dimensions that their measurability may be examined (see Chapter 2) and their special attributes identified. It must be entirely clear whether they are directly observable or matters of inference only, whether items of recollection or of present feeling, whether aspects of a behavior or a construct about behavior, etc. Furthermore, the definition should be in such terms that appropriate procedures and forms for measurement can be devised directly from the definition. This means that the definition must be in terms of behavior and not abstractions.

2. Because behavioral measurement too frequently occurs under artificial circumstances, the concept of the simulated task or prototype behavior has developed. Obviously, if the phenomenon to be measured is writing ability, then the pupil should be measured while he is writing. If this is not possible, he may be measured doing things that resemble or simulate his ordinary writing behaviors or that embody the essential actions of any writing behavior. The more the behavior during measurement approximates the actual behavior for which a measure is desired, the more valid is that procedure. If even a simulated behavior is impossible, then responses must be contrived for measurement that are correlated to the highest degree possible with the behavioral phenomenon in question. A good mathematics test may have high validity on the basis of employing simulated tasks. Even the best personality test has to base its claim to validity on the second approach, which we will call indirect measurement. (For examples of simulated task tests see Chapter 13, pages 347-350, in particular.)

3. A third variable in validity is the extent to which the measuring procedure appraises in proper proportion and perspective all the significant dimensions of the phenomenon being measured. For example, reading involves rate and comprehension for varied materials and for varied purposes. A procedure for measuring reading ability that involved comprehension, one type of material and reading for one purpose only, memorization say, would be less valid than a procedure that measured rate also, used three different materials (text-book, fiction, and newspaper) and entailed reading for purposes of searching and skimming as well as for memorizing. If one or more of the dimensions consist of components or parts and these are in any way extensive, the measur-

³ Statistical expressions of validity are examined in Chapter 8, page 186.

ing procedure must engage a sufficiently large and sufficiently representative sample of these components. This is the major variable in the validity of vocabulary tests where extensiveness of vocabulary is the chief dimension.

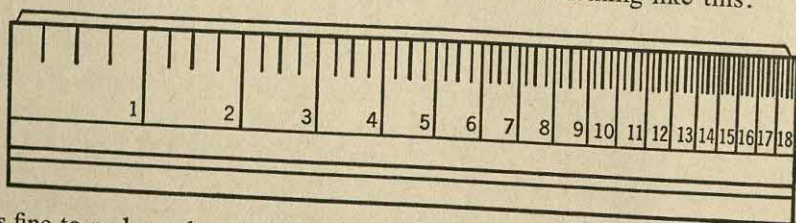
4. A fourth consideration in validity involves the extent to which the results obtained by the procedure are produced by the dimensions purportedly being measured and not by other unpredicted variables. Suppose that a general science test had many unnecessarily long sentences and employed many unusual words. Very good readers would make higher scores than average readers, even if they had an equivalent knowledge of general science. Thus, this test would be invalid to the degree that differences in reading ability affected the results. Reading is a very usual unpredicted variable that tends to invalidate measuring procedures to some degree. Other common ones are test anxiety, differences in intelligence and motivation among pupils, and bias on the part of the examiner.

Each type of measuring procedure has, of course, certain unique problems relative to validity and for each there are specific means of increasing validity. These will be presented when each procedure is discussed later in this chapter. Ways of estimating and expressing the validity of procedures are treated in Chapter 8, page 186.

RELIABILITY

The second criterion of adequacy in a measuring procedure is reliability. Reliability means that the procedure measures consistently and uniformly over the duration of the procedure and in a reapplication of the procedure. A steel ruler is a good example of a thoroughly reliable measuring device. No matter when or where you use it or how you hold it, its inches are the same length and the same distance is measured to be the same time after time.⁴

A "fisherman's" ruler and a rubber ruler are examples of unreliable measuring devices. The "fisherman's" ruler looks something like this:



It is fine to make a short fish long enough to talk about, but not very good for measuring anything else. The trouble with it is that its inches are not of uniform length. If you measure at the short inch end, and inch could be as long as eight inches. Thus, the *same distance* will appear to vary as to length according to what part of the ruler you use. A rubber ruler, on the other hand, may

⁴ Slight deviations in readings due to varied positions of the eye, differences in lighting, and the expansion-contraction of the metal do occur. They are so slight, however, as to be negligible for most purposes.

have inches of equal length but all of it or any part of it may be stretched. Hence, depending upon how tight or loosely the ruler is held during measurement, distances apparently become shorter and longer.

Analogous to the "fisherman's" ruler is a test uneven in quality (typography, directions, vocabulary). The student who happens to know the answers to the good portions of the test is favored, whereas the student whose knowledge relates mostly to the poor parts is hampered. The rubber ruler test is one in which many questions are ambiguous or whose directions are subject to varied interpretation. In January a pupil may define the words one way and interpret directions accordingly. In June he has forgotten his January definitions and interpretations and makes new ones. It is hardly a surprise that his score is different in June.

In educational measuring devices the intent, of course, is to approach the degree of reliability possessed by the steel ruler. The requirements for maximum reliability in each basic procedure are discussed in subsequent chapters; but by and large, length and clarity are the critical variables for reliability. As procedures are more extensive, they tend to sample more adequately and the relative effect of a few bad items is diminished. Clarity of expression prevents the ambiguity and misinterpretation which, along with deficient sampling, constitute the more important sources of unreliability.

Several statistical procedures are available for determining the degree of reliability possessed by a given procedure. Degree of reliability is expressed mathematically as a correlation coefficient and is called a *coefficient of reliability*. The size of coefficients for published standardized tests tend to range from .70 to .95, with the most reliable achievement batteries having coefficients of .90 to .95. The significance of coefficients of different sizes together with techniques for deriving them are explained in Chapter 8, pages 172-181.

EFFICIENCY

Efficiency, as you know, means that the most output is accomplished with the least input; and you know also that it is a criterion for many human activities other than measurement. With reference to measurement, efficiency means that a given procedure provides the needed measurement symbols in minimum time and with minimum expense and energy when compared with other possible procedures. Even more than validity and reliability, is it a function of specific types of procedure; you are referred to the treatments of these procedures for means of increasing their efficiency.

It should go without saying that time and energy saved in measurement *at the expense of reliability and/or validity* is actually time wasted. However, the relationship between reliability and validity and the duration of a measurement procedure is curvilinear and, after a point, large increases in the length of a test or other procedure result in only slight increases in reliability.⁵

⁵ See page 114 in Chapter 6 for an analysis of the relationship between length of a test and its reliability.

Moreover, the most careful and time-consuming revision of an already reasonably valid test can be expected to yield only slight gains in the test's validity. The curve of Figure 2 suggests the relationship likely to exist between reliability or validity and time or energy devoted to a measurement.

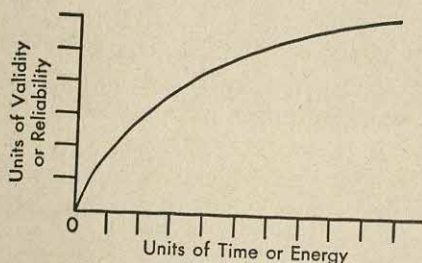


Figure 2. Probable relationship between the reliability or validity of a measurement procedure and the time and energy devoted to the procedure. (This curve is illustrative only and has no precise mathematical significance.)

distances on highway signs need only to use odometers whose finest reading is one tenth of a *mile*. In general, the greatest degree of precision is necessary for research purposes, the next for decisions and predictions about individual pupils, and the least for decisions and predictions about groups of pupils.

Some Technical Terminology of Measuring Procedures ⁶

As behavioral measurement has been developed by teachers and psychologists over the last fifty years, a number of terms have taken on special, technical meaning for the procedures of measurement. Several of these terms are significant for all types of procedures and need to be understood if descriptions of the procedures are to be meaningful.

Instrument. Any tangible device used to assign measurement symbols to phenomena. A test, rating form, record sheet, check list, test battery, etc., are all called *instruments*.

Test. A special type of measuring instrument. Its general characteristic is that it forces responses from a pupil and the responses are considered to be indicative of the pupil's skill, knowledge, attitudes, etc. True-false tests, essay examinations, attitude scales, short-answer tests, mid-terms and finals, and a personality inventory are all technically to be called *tests*.

Item. A part of a test that elicits a specific response. Examples of an *item* are a multiple-choice question, a problem on a mathematics test, a true-false question, and a picture in a projective personality test.

Standardized. The instrument (usually a test) so designated has previously been administered to a population with known characteristics. A

⁶ In the Appendix A is an extensive glossary of terms connected with educational measurement and evaluation.

pupil's score may then be interpreted with reference to the scores made by the group of pupils upon whom the test was standardized. This group is called the "standardization sample." For example, the *Davis-Eells Games* were administered to approximately 3,000 pupils per grade in order to determine the significance of different scores.⁷

Norm. The score on a standardized test that is typical of a given age or grade of pupils or of some segment of the age or grade population. In the test cited above, a score of 41 is the average score for pupils nine and a half to ten years old. A score of 51 represents the 91 percentile rank among this age of children. Both the average score and the 91st percentile score are *norms*. The derivation of norms involves statistical techniques and these are explained in Chapter 7.

Bias. The tendency of an instrument to favor a certain outcome or to discriminate for or against pupils with certain characteristics. Most paper and pencil tests designed for use with groups are *biased* in favor of good readers. That is, if two pupils have equivalent knowledge, the one who is the better reader is likely to make the higher score. Many questionnaires distributed by political organizations are so constructed that the replies inevitably favor the policies of the political organization.

Criterion. That with which a procedure and/or instrument is compared to determine its validity. The ratings given by clinical psychologists and psychiatrists to a group of people often constitute the *criterion* of validity for a personality test. In intelligence testing, the Revised Stanford-Binet Scale (Appendix B) for many years has been the *criterion* for intelligence tests. The degree to which a newly devised test produced results similar to those produced by the Binet was offered as evidence of the new test's validity.

Scoring. Checking test responses against a key to determine their correctness/incorrectness or to categorize them in some other way; assigning numbers or other symbols to pupil products or free expressions so as to rate them.

Summary

The function of behavioral measuring procedures is to assign the proper measurement symbols, usually numerals, to dimensions of behavioral phenomena. To accomplish this, the procedures employ *standard stimulations*, elicit *standard differential responses*, use *standard analysis systems*, and engage in *sampling*. All procedures do not employ all the devices but all involve sampling. Samples should be reasonably large and representative both categorically and temporally for accurate measurement.

Procedures in use may be classified as observation, product analysis, free response tests, and guided response tests. Three criteria are applicable to determining their effectiveness: validity, reliability, and efficiency. Validity usually is defined as the capability of a procedure to measure what it purports

⁷ Davis, Allison, and Eells, Kenneth, *Davis-Eells Games*. New York: World Book Co.

to measure. Reliability means that the procedure measures consistently and uniformly over the duration of the procedure and in a reapplication of the procedure. Efficiency, of course, asks that a given procedure provide the needed measurement symbols in minimum time and with minimum expense and energy.

A number of terms are employed in special technical ways in educational measurement. Among them are instrument, test, item, standardized, norm, bias, criterion, and scoring.

EXERCISES

1. List the several aspects of behavioral phenomena that make their precise measurement difficult.
2. Select any single line in this book and count the number of times each letter of the alphabet occurs. Repeat this process for another line and compare the results of the two counts. What does this imply for sampling? Continue to count the frequency of letters in lines until *you are sure* that you know the relative frequency of letters *throughout the book*. How many lines did you use? Why were you sure a given number of lines was enough? Is the relative frequency of letters in this book the same as in English writing generally?
3. Using a specific test, state how its validity, reliability, and efficiency are interrelated.

BIBLIOGRAPHY

1. Anastasi, Anne, "Concept of Validity in the Interpretation of Test Scores," *Educational and Psychological Measurement*, 10:67-78, 1950.
2. ———, *Psychological Testing*. New York: The Macmillan Co., 1954.
3. Comrey, A. L., "Mental Testing and the Logic of Measurement," *Educational and Psychological Measurement*, 11:323-334, 1951.
4. Cronbach, Lee J., *Essentials of Psychological Testing*. New York: Harper & Bros., 1949.
5. Gaylord, R. H., and Stunkel, E. R., "Validity and the Criterion," *Educational and Psychological Measurement*, 14:294-300, 1954.
6. Gulliksen, H., *Theory of Mental Tests*. New York: John Wiley & Sons, Inc., 1950.
7. Lindquist, E. F. (ed.), *Educational Measurement*; ACE, Washington, D.C., 1951.
8. Mosier, C. I., "Critical Examination of the Concepts of Face Validity," *Educational and Psychological Measurement*, 7:195-205, 1947.
9. Stuit, D. B., "Effect of the Nature of the Criterion upon the Validity of Aptitude Tests," *Educational and Psychological Measurement*, 7:671-676, 1947.
10. Thomas, L. G., "Mental Tests as Instruments of Science," *Psychological Monographs*, LIV (1942), No. 3.

11. Travers, R. M. W., "Rational Hypotheses in the Construction of Tests," *Educational and Psychological Measurement*, 11:128-137, 1951.
12. Walter, W. Grey, *The Living Brain*. New York: W. W. Norton & Co., Inc., 1953.

CHAPTER 4

OBSERVATION

The most primitive of man's procedures for measurement is simple observation. The nomad hunter looked at tracks, listened to animal cries, felt the breeze against his cheek, watched the moon and stars, all to determine what game were present, what the weather was likely to be, and what season was approaching. Yet, two of the most important of civilized man's vocations, medicine and psychology, employ observation as their principal technique of measurement. For all his meter readings and laboratory analysis, it is largely what the physician's eyes see, and what his probing hands feel, that tell him what your illness is and how serious it is. And, although the psychological clinician may use a Rorschach test and even an electroencephalogram, his diagnoses of the neurotic's phobias and the psychotic's delusions depend heavily upon what he observes of the patients' talk, gestures, and facial expressions.

We shall define observation as *measurement without instruments*, or, if instruments be used, they affect the measurer and not the measured. In observation procedures, the measurer applies his senses directly to the phenomenon being measured. He looks at the behavior (studying), he does not look at a score on a test that is indicative of the behavior (a score on a test of study habits). Because, in observation, we apprehend the dimensions of phenomena directly, observation, of all procedures, is most susceptible to errors of faulty perception and bias. For this reason, it is a primitive technique often despised by the scientist for its unreliability. But, just because dimensions are apprehended directly, all types of sensory data can be handled simultaneously and in proper relationship. Moreover, the pattern of a phenomenon can be seen as well as its elements. Thus, observation can be a highly civilized procedure, valued for its validity by the same scientist who decries its unreliability.

In education, observation is the most widely used of all measurement procedures. Principally, this is because many behavioral phenomena may not be assessed validly by any other procedure. In succeeding paragraphs we will explain some of the factors that affect the usefulness of observations, describe several observation techniques, state the forms of measurement that observations may be expected to yield, and suggest the types of educational phenomena for which observation is a legitimate measuring procedure.

Standard Analysis and Adequate Sampling in Observation

In observation, an attempt is made to appraise whatever happens, as it happens. Consequently, neither standard stimulations nor standard responses may be utilized to make the measurements more accurate. The observer must try to make his observations competent just through proper use of a standard analysis system and through adequate sampling.

To insure that the same measurement symbol is assigned to phenomena having the same characteristics, you may recall that it is necessary to have a standardized way of reacting to phenomena. In observation, this standardized reaction is especially difficult to attain. For one thing, what is to be observed is entirely a junction of the observer. There is no test or other instrument present to force a given desired behavior. What behaviors are occurring is largely happenchance, and what the observer sees and hears of what does occur is entirely up to him. For example, a teacher may wish to observe Bill's work habits during the fourth period. Bill may or may not work during the fourth period. If he does work, it is still up to the teacher to keep his attention on Bill, to attend to Bill's actions rather than Bill's appearance, and to watch Bill's work-oriented actions and not his socially oriented actions. Then, the teacher may wish to observe Nancy's work habits, so as to compare them with Bill's. Again, the same disciplined perception is required, and, in addition, the observation of Nancy must be like that of Bill.

A second deterrent to standard analysis in observation is the fact that the observer necessarily is aware of his own feelings as well as of external happenings. It is as if a thermometer were at the same time a wind gauge yet seemed only to measure the temperature. As the wind increased and decreased, the thermometer would *show* changes in temperature even though the *actual* temperature perhaps remained constant. Just so, a teacher needs to observe and rate the "citizenship" of pupils but finds it difficult to keep his stomach, muscles, and heart out of the picture. When he is well-fed, rested, and happy, the pupils may be good citizens. When he is hungry, or dyspeptic, tired or sad, they are more apt to be considered hellions. Yet *the pupils* may not have changed at all.

The problem of sampling is less complex for observation than for some other procedures, but it is no less critical. So far as temporal sampling is concerned, it is known that children and youth fluctuate widely over relatively short periods of time in their motivation, achievement, interests, and nearly everything of educational significance. Uniformity there is, but it is the uniformity of a mountain range rather than that of a plain. Yet, the teacher frequently is called upon to evaluate motivation, achievement, and interest as if they were always the same for a given pupil. It is little wonder that his judgments are sometimes in error when, as too often happens, they are based on only a few observations of the pupil.

Categorical sampling (representing properly the different elements of the

phenomenon) is even more of a problem in observation. We know that a child's progress in speaking properly involves his improvement in enunciation, rate, tone, vocabulary, usage, etc. If the speech of pupils in a ninth-grade class is to be measured accurately, the same attention should be given to each pupil with regard to each of the dimensions. If some are heard only in situations that involve easily pronounced words and some in situations involving difficult words, if some are observed in recitations where there is little opportunity for tone variation and so on, categorical sampling obviously is inadequate.

Some devices are presented shortly (pages 52-59) that will facilitate standard analysis and representative sampling. However, you will find that the observer's skill and concentration are more critical than his technique. It is necessary to know the significance of standard analysis and representative sampling, and then consciously to practice them, if you wish to be a competent observer.

Forms of Measurement Appropriate to Observation

Historically, observation has been a means to measurement by *description*, *classification*, and *ranking*, but rarely to scale measurement. Teachers in both ancient and modern times have watched pupils and have labeled them good, fair, or poor in achievement; lazy or industrious in study; and well-behaved or unruly in deportment. Then, again, teachers have listened to speeches and ranked them 1, 2, 3. They have designated Henry as third in his class, and Marie as seventh. But, with a few exceptions, they have not said a pupil is eleven years and six months (a scale number) on the basis of observation alone. Nor have they assigned him an IQ (another scale score) nor indicated how many seconds it took him to run a race from simple observation only.

Similarly, except where counting is involved current observation procedures may be expected to yield only descriptive and classificatory symbols and rank numbers. The descriptions produced by observations are discussed under check lists and anecdotal forms (pages 52-54). Classification symbols utilized in observation range from the simple dichotomies of satisfactory—unsatisfactory, obedient—disobedient, to the many valued classificatory systems of certain rating scales (see pages 56-57). Where numbers that do not indicate a count are recorded as the result of observation, they usually represent verbally defined classifications or rank within a group. It is possible to derive scale numbers from observations, but the techniques are so time-consuming as to be justified only in research. Hence, the significance and accuracy of observational measurements usually are limited by the significance and precision of descriptive, classificatory and rank symbols.

Evaluation as a Direct Product of Observation

Observations frequently produce evaluations rather than measurements. When, as a result of observation, a tenth-grade pupil is given a grade of *B*, a beginner's performance on the piano is awarded a certificate of merit, or an

emotionally disturbed child is placed in a mental institution, an evaluation has been made, not a measurement. In the introduction to Section I, we differentiated between characterizing the status of phenomena (measurement) and judging the value of phenomena (evaluation). The latter, you may recall, involves assigning a value symbol to something, usually by comparing its status with some standard of value. Such a comparison between status and standard was made in the case of the grade, the certificate of merit, and the institutional placement. Probably, the teacher had in mind several levels of achievement for tenth graders and he thought that the pupil's achievement matched the level that was worth a *B*. The piano teacher knew about what to expect of beginners. The beginner appeared to live up to expectation and, hence, was judged to deserve the certificate that symbolized that quality of piano playing. And the psychiatrist who judged that the child needed confinement and treatment possibly did so because the child's behavior seemed to be similar to that of a classification of symptoms called schizophrenia and hence was evaluated as requiring hospital treatment.

Evaluations based on observation are often subjective. The standard used by the tenth-grade teacher was in his own mind, as was that of the music teacher and the psychiatrist. Now, as we shall see in Chapter 9, evaluations tend to lose validity when they are made subjectively. Moreover, when the evaluation of a phenomenon is the *only* thing expressed, it is not possible to verify the measurement on which the evaluation is based or even to ascertain that any separate act of measurement occurred. Thus, the tendency of observation to produce evaluations directly and immediately may produce errors in measurement, in evaluation, or in both.

To guard against such errors it usually is advisable first to observe and characterize the actual status of the thing in question and then to compare this status with a standard, keeping both steps separate and explicit. The standard to be applied should, moreover, be written down or given objective form in some way. If, for some reason, the measurement aspect of observation may not be explicit, it may at least be kept separate from the evaluation. This requires that you, as an observer, must know when you are *seeing* the characteristics of a pupil, and when you are *judging* what you have seen. It requires, moreover, that you must consciously perform the measurement first and the evaluation second.

Evaluations that derive immediately from observation often are called *ratings*, and the devices used for such evaluations usually are called *rating scales*. These are discussed on page 55. The primary functions of *evaluative* rating scales (some are simply measuring devices) are to give more objectivity and clarity to standards and to insure more systematic comparisons between phenomena and standard. Thus, their use serves to increase the validity of observational evaluations.

Devices Used in Observation

A number of verbal and graphic devices are used in observation. In general, these are designed to insure standard analysis and adequate sampling, to mitigate against too quick and too subjective evaluations, and to predetermine the forms in which measurements are to be expressed.

CHECK LISTS

A check list is any sort of device, simple or complicated, which contains the items to be observed and, perhaps, space for number or short verbal entries. The check list in Figure 3 is an example of one that a physical education instructor might use. Other check lists provide for an indication of the presence or absence of certain factors. These are employed frequently in surveys of school plants and facilities. One that could be applied to the audiovisual services of a school might contain such items as: Central storage of equipment _____, Storage in classrooms _____, Teachers operate _____, Students operate _____. Checks would be placed by the items that obtain for a given school.

	Name

	Grade-Age-Date
1. Type of activity	
2. Interest	
3. Effort	
4. Co-ordination	
5. Posture	
6. Skill in activity	
7. Sportmanship	
8. Other factors	
.....	

Figure 3. Check list for observation in physical education.

Check lists sometimes are called observation schedules, particularly when they are lengthy. They are employed extensively in Homemaking, in Industrial Arts and Agriculture, in Physical Education, in Art, and in all other school areas where observation is an important procedure of measurement. They help make observations more reliable by stimulating the observer to look for the same factors or dimensions each time he observes. They do not eliminate bias, they do not insure adequate time sampling nor uniform recording.

Few check lists are available commercially, and in most subjects, for most purposes, they are not especially desirable. Published check lists are likely to

be no more reliable than self-constructed ones, and they are likely to be less relevant to your measurement task. To design one, enumerate the dimensions you wish to observe, define them clearly, eliminate those that are vague or repetitive, arrange them on a sheet of paper in whatever manner is most convenient for observing and recording, include space for identifying data, and try out the form. When it seems workable, duplicate a small supply. When you have used up this pilot batch, revise the check list on the basis of your experience, and run off the quantity you will need.

ANECDOTAL FORMS

An anecdotal record is a check list that provides space for much writing, and, as a rule, provides for less breakdown of dimensions than a check list. Often the forms are used over a period of time and are meant to be cumulative. Their significance is primarily to minimize the use of high-level abstractions in recording and to obtain instead an "operational description" of behavior. An anecdotal record designed to serve much the same purpose as the physical education check list in Figure 3 is shown in Figure 4.

Purely as paper-and-pencil aids, anecdotal forms do little more than do

_____ Name	
<i>Directions</i>	
In the space provided, record observations that bear on the individual's physical development and social development. <i>Do not evaluate, but describe.</i> Avoid vague words such as good, strong, shy, etc. Enter statements of what happened, or what you saw, as "Did three push-ups, and couldn't do any more," "Cried and started fighting when he was called out." <i>Date each entry.</i>	
Physical Development:	
Social Development:	

Figure 4. Anecdotal record for physical education.

blank sheets of paper to insure validity and reliability in observation. As a viewpoint about recording observations and as a means to enforcing that viewpoint, they have considerable merit. The viewpoint is that observation records are valid and reliable to the extent that they actually reproduce whatever was observed. The best anecdotal record is a sound motion picture. In practice this ideal is not achieved. What can be achieved is a relatively unambiguous record of things that a teacher noticed and thought important about a pupil's behavior. Proper use of anecdotal recording tends to make for more adequate temporal sampling, but categorical sampling often suffers since the observer is not recurrently reminded of what to observe.

Anecdotal recording is used most frequently by elementary teachers and in courses where citizenship or social development is an express goal. In addition, counselors, school psychologists, social workers, and supervisors make extensive use of anecdotal records. As with check lists, there are relatively few published forms and self-designed anecdotal forms are generally more satisfactory. The construction of the form is even less difficult than the construction of a check list, but to record properly is somewhat more difficult. Proper anecdotal recording is characterized by the following:

1. What is written down is what was seen or heard. Inferences, guesses, assumptions, are omitted unless they are clearly labeled as inferences, guesses, or assumptions.
2. The observer has determined what aspects of behavior are related to the dimension being appraised. He observes these only and records these only.
3. If the record is to be cumulative, a plan of periodic observation and recording is established and adhered to.
4. Words and phrases are used whose meaning is clear, and, so far as possible, unequivocal.
5. Words and phrases are employed that are definable in terms of things rather than other words. Concrete statements are preferred to abstract ones. For example, "He became pale and his hands trembled," not "He was disturbed."
6. Words and phrases that have strong emotional connotations are avoided, i.e., love, hate, insolent, courteous, loyal, dishonest, etc.
7. Words and phrases are avoided which express the observer's judgment, or his opinion, and not just his perception. Among the frequently encountered "judgmental" terms that should be avoided are these:

- | | |
|-----------------|----------------|
| a. well-behaved | e. industrious |
| b. delinquent | f. nervous |
| c. aggressive | g. happy |
| d. didn't try | |

RATING SCALES

As observational aids, rating scales are at the opposite end of the spectrum from anecdotal forms. Recording is brief and formalized but much is printed

on the form, whereas anecdotal forms provide for extensive and informal recording and contain far more blank space than printing. The types of rating scale are legion and the example items shown in Figure 5 can only suggest their myriad forms.

Measurement Scales and Evaluative Scales. Of the rating scale items shown in Figure 5, *B*, *D*, and *F* are primarily devices for recording status (measurement) while *A*, *C*, and *G* seem to be means of expressing judgments (evaluation). In *B*, *D*, and *F* you may observe that the scale merely covers the possibilities of the dimension in question. Any high or low value placed on a given rating is a matter of implication or for later determination. In *A*, *C*, and *G*, on the other hand, the observer clearly is expected to evaluate the pupil on the dimensions in question. Not only is he to observe what the pupil's condition *is*, but he also is to compare this status with a standard and thus *judge* the pupil: "satisfactory" in study habits, a "poor" performer on the playground, possessed of the "best possible" attitude toward school.

Scale Including an Evaluative Standard. Item *E* illustrates a rating scale that is a *measurement* device but has an *evaluative standard* superimposed upon it. Where the standard and the measurement scale are related properly, the device has great merit. Subjectivity in evaluation is eliminated for the observer. He need concentrate only on finding the point on the scale that best expresses a pupil's status (in posture in Example *E*) and the evaluation is made automatically. The propriety of the evaluation is then not a function of the observer, but of the rationale underlying the scale. In the case of Example *E*, physicians' opinions as to the relationship between different postures and health presumably is the basis for the evaluative scale and evaluations are valid to the extent that these opinions are valid.

Category Scales. Rating scales sometimes ask for phenomena to be assigned to categories and sometimes to be placed along a continuum. Examples *A*, *B*, *C*, and *E* are categorical rating scales. In each instance a pupil is to be given a definite classification, *S* or *U*; Always, Usually, or Seldom; etc. In such ratings, only gross differences between pupils are appraised, and tacitly it is assumed that all the pupils assigned a given categorical rating are alike in the dimension being appraised.

Continuum Scales. Continuum scales are exemplified by items *D*, *F*, and *G*, in Figure 5. Example *F* is a pure continuum. A line is shown to represent the variation between two extremes. A pupil's status is to be indicated by checking any point on the line, and, in theory, each pupil might be checked at a different point, and thus the *unique* status of each pupil is retained in the rating. Modified continuums are shown in *D* and *G*. The rationale for these two scales is the same as for *F*; extremities of status are established and degrees of variation between the extremities are represented. Rather than unbroken lines, however, numbers and phrases are shown at intervals along the line, and the observer's task is to choose the appropriate number or phrase. These two continuum scales are analogous to a ruler that contains inches, but no fractions of inches. Distance measured by the ruler may be read to the nearest

inch. Variation as to persistence may be marked to the nearest phrase, or variation as to attitude toward school to the nearest number. If the observer wishes, he may interpolate between phrases and numbers, just as with a ruler you may estimate 3.5 inches, when only 3- and 4-inch intervals are marked off.

Graphic or Descriptive Scales. Items *D* and *F* in Figure 5 illustrate two other possible characteristics of rating scales. In *D* each phrase is given a number value and, though a pupil is rated according to his approximation to a phrase, his rating is recorded as a number. Such use of corresponding numbers and phrases is widespread. These scales usually are called graphic rating scales.

(A)

(Write a letter)

Rate the child on each trait listed as satisfactory or unsatisfactory by putting *S* or *U* in the space provided.

_____ Honesty

_____ Obedience

_____ Neatness

_____ Study Habits

(B)

(Check a column)

Always

Usually

Seldom

Keeps his desk, books,
and other materials
clean and neat.

(C)

(Write a number)

Evaluate the pupil's behavior in regard to each factor by placing 1, 2, or 3 in the box by the factor.

1. like the best pupil's behavior

2. like the average pupil's behavior

3. like the poorest pupil's behavior

Playground activity ☐Going to and from school ☐Attitude toward teachers ☐

(D)

(Choose a phrase)

Is he easily discouraged or is he persistent?

Melts
before slight
obstacles or
objections
(5)

Gives up
before
adequate
trial
(3)

Gives
everything
a fair
trial
(1)

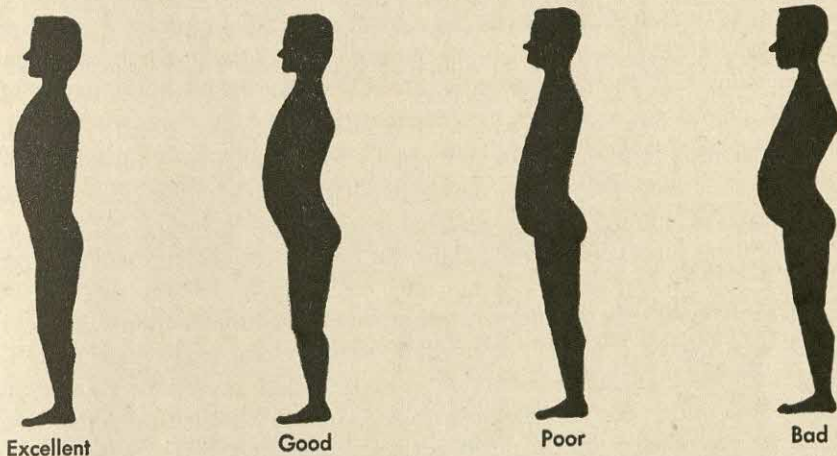
Persists
until
convinced
of a mistake
(2)

Never
gives in,
obstinate
(4)

(E)

(Choose a picture) *

POSTURE STANDARDS
Intermediate-Type Boys



(F)

(Check a point on a line)

Moral
Standards

Extremely
moral,
hyperactive
conscience

No apparent
self-standards,
always follows
the crowd or
expediency

(G)

(Check a number on a line)

Attitude
toward
school

1	2	3	4	5	6	7
Worst possible		Average		Best possible		

Figure 5. Types of rating scale items.

* Adapted from Arman Klein and Leah C. Thomas, *Posture Exercises*, Children's Bureau Publication, 165. Washington, D.C., U.S. Department of Health, Education and Welfare, 1926, p. 12.

Scales With Unequal Intervals. In Item G you may notice that numbers are unevenly spaced; 4, the center number, represents an interval four times as great as 1 or 7, the numbers at the extremes. The use of unequal intervals has both a psychological and a statistical basis. It is known that it is more difficult to make valid distinctions among the large group of pupils we call average or

normal than it is to note differences among those children who deviate from the average to an extreme degree. A *G*-type scale compensates for this condition by spreading the size of the middle intervals and constricting the size of the extremity intervals. From a statistical point of view, a *G*-type scale forces the ratings of large numbers of pupils to approximate a normal distribution (see Chapter 8, pages 161-169) and thus to approximate the assumed real distribution for many human dimensions.

Measurement Symbols Derived from Scales. With the use of continuum rating scales, it is possible to obtain numbers from which rank or order symbols may be derived. Obviously, if one pupil is given a rating of 3 and two others ratings of 4 and 5, each is asserted to have more or less of a dimension than the others, and consequently the rank of each of the three is apparent from their ratings.

It is not possible as a rule to attribute scale position to the numbers of rating scales, even though they are called "scales." You may recall (Chapter 1, pages 6-8) that the numbers of a valid scale represent known and usually equal intervals, and that the numbers count from a zero or other fixed and known reference point. In our observational rating "scales" it is apparent that these conditions are not met.¹ Whether or not line subdivisions and/or numbers represent equivalent increments of difference is entirely a matter of conjecture. Moreover neither the center nor the ends of the usual rating scale may be considered to have fixed values.

Number of Categories or Subdivisions in Scales. The number of categories or of subdivisions optimum for a rating scale is indeterminate. The usual number of intervals is five, but there is no rational justification for this number. Wrightstone, reviewing research in rating methods, asserts that seven is an optimum number for rating human traits (9:962). Rating scales with more than ten units are unusual, and a two-unit scale (a dichotomous rating) is, of course, the bottom limit. The principle to be followed in designing a rating scale is that the number of scale intervals should approximate the number of clearly discernible differences in the dimension being appraised. For measurement tasks that require great precision a greater number of scale units may be necessary. For tasks requiring less precision, fewer intervals are permissible.

Sources of Scales. A large number of rating scales are published but, as with check lists and anecdotal forms, they are not subject to standardization. Many of the scales designed each year for research purposes are described in

¹ In research situations, rating scales have been and may be developed with approximately equal intervals and with somewhat fixed points of reference. Scale numbers can be derived from such "scales," but they are to be interpreted with caution. In ordinary school situations, rating "scales" are scales only in the figurative sense of the term. Methods by which approximately equal intervals and fixed reference points may be established are described in L. L. Thurstone, "The Method of Paired Comparisons for Social Values," *Journal of Abnormal and Social Psychology*, 21:384-400, 1927 and J. P. Guilford, *Psychometric Methods*, New York: McGraw-Hill Book Co., Inc., 1936.

professional journals and constitute a good source of ideas. Buross' *Mental Measurements Yearbooks*, publishers' catalogues and reviews, and advertisements in professional periodicals provide listings of the published scales.

Applications of Scales. Rating scales find their greatest use, of course, in areas where measurement must rely largely on observational methods. Hence, they are employed extensively in the appraisal of personality, social behavior, and teaching competence. Curriculum evaluation frequently utilizes rating scales, as do surveys of school housing and facilities. Report cards (see Chapter 9) are in effect evaluative rating scales whose ratings are based on many observations and/or testings in broad categories of achievement.

Criteria for Rating Scales. Certain generalizations are appropriate to the proper design and efficient use of rating scales.

1. The dimensions to be rated need to be very clearly defined in terms of what is to be observed when the dimension is rated. The dimensions, moreover, need to be as distinct as possible, each one being observable and measurable by itself and not overlapping with another. Only as a dimension can be clearly and discretely defined in operational terms can it be rated with any validity. Thus, *enunciation* is a dimension of speech that can be rated effectively, while ratings of *interest* are nearly always suspect.

2. If several dimensions are to be rated for a given pupil, it is probable that each rating will be influenced by the others and in the same direction. This is called the "halo" effect. If an initial rating is high, others will tend to be high; if low, the others will tend to be depressed. Another way of viewing "halo" effect is that the observer has a feeling about a person as a whole, and he then proceeds to record ratings that will justify this feeling. To hedge against this type of error, you need to concentrate on each separate dimension you rate and consciously "shut out" your feelings about other dimensions or the person as a whole. Frequently rating scales are devised with the right extreme of a scale sometimes having a low value and sometimes a high value just to hamper the operation of a "halo" effect.

3. Some observers are found to rate low and others to rate high, no matter who or what is being rated. If such proneness to over- or underrate is known, the observer must take steps to correct it or provide for an automatic correction after he has made his ratings. Lack of experience with the full range of variation in the item being rated often is a cause of consistently high or low ratings.

4. Other observers tend to avoid the extremes in their ratings. They unconsciously use middle or average ratings unless the deviation is so great as to force the use of a higher or lower index. And, even in this case, they will still record ratings closer to the center than an unbiased observer would.

Validity and Reliability in Observation

Preceding paragraphs have dealt with the nature and limitations of observation as a measuring procedure, and with the devices and techniques that

may facilitate observation. In this presentation a great many factors necessarily have been stated that bear on the validity and reliability of particular aspects of observation or of its use for particular purposes. Now, in conclusion, we would like to present some validity-reliability maxims for observational techniques as a whole.

Use Where and When Legitimate. Observation is justified when testing or product analysis is too time-consuming, too expensive, or when the dimensions to be measured cannot be measured validly except by observation. As we have observed, reliability is likely to be lower for observation than for other procedures and, consequently, it is justified only when more reliable procedures are not equal to or available for the task. The school subjects and areas in which observation is most likely to be an essential procedure of measurement are Art, Music, Homemaking, Physical Education, Speech and Dramatics, and citizenship, social adjustment, study habits and the like.

Use Appropriate Paper-and-Pencil Devices. The function of any observational device, check list, anecdotal record, or rating scale, is simply to insure the conditions of standard analysis and adequate sampling that observation requires and, in addition, to mitigate against some of the human errors inherent in observation. Use of the device most appropriate to the given observation task is mandatory. Appropriateness of a device is a function of the form of measurement desired, of the purpose of the observation, and of the phenomenon and/or dimensions being observed. Obviously, a check list or an anecdotal form is needed for descriptive data and a rating scale of some sort for classification or ranking. The purpose of individual diagnosis is better served by an anecdotal form on a check list; and that of comparing pupils, by a rating form. In general, phenomena with many complex and ill-defined dimensions are better approached through anecdotal records, while check lists and ratings may be used for those with simple and well-defined dimensions.

Record Quickly. It is axiomatic that good observation records or ratings are made during the observation or immediately thereafter. Only by such quick recording can you insure that the record is *strictly a function of what was observed*. As time elapses, various types of distortion are likely to occur.² Detail will be forgotten. Intervening experiences will interdict or become confused with the conversation. Remembrance of what was observed will become progressively more like the stereotype it approximates. Strong impressions will become stronger, and weak ones weaker.

If the keeping of a record *during* a period of observation may invalidate the observation, notes must be made immediately after. Even a wait of thirty minutes or a few hours is likely to invalidate the record. As a rule, if a person is likely to construe that your observation or its outcome has any effect on his

² The distortions that occur between an event and its recollection have been studied most thoroughly by "field" psychologists. Their findings and generalizations are reviewed in Ernest Hilgard, *Theories of Learning*, New York, Appleton-Century-Crofts, Inc., 1948.

reputation or prestige, or even on your opinion of him, recording during the observation is improper. You may proceed to record during the observation only when you are sure that the fact of your recording is not an important stimulus to the individual.

Guard Against Bias. The existence of some preconception or feeling set toward the person or thing observed is highly probable. It is advisable then to assume that such bias exists, to determine what it is, and, consciously, to try to keep it from affecting the description or rating you record. If the bias is extreme, you should disqualify yourself as a competent observer, just as a judge does when he has a special interest in a case being tried. One way of verifying your control of your bias is to record as objectively as possible, and at great length, and then to ask an impartial yet qualified third person to make an independent rating on the basis of your record. Unless he agrees essentially with your rating, it is well to question your own rating.

Base Evaluations on Several Observations if Possible. Obviously, a judgment based on one measurement is less sure than one based on several measurements, even when the measures are highly reliable. When measures are less reliable, use of several measures as a basis for evaluation is even more important. As we have seen, observation has the maximum likelihood for unreliability of all the procedures of measurement. The number of observations may be increased by pooling the observations of several observers as well as by repeating your own observations. If several observers are used, it is well to assure that all are competent and that all observe the same dimensions in the same way.

Summary

A great deal of educational measurement is done with little or no assistance from instruments. "Observation" is a collective term for the various methods of this largely unaided measurement. In observation an attempt is made to appraise whatever happens, as it happens. Consequently, only standard analysis and sampling are employed, not standard stimulations nor standard responses. What is observed is necessarily a function of the observer and thus inattention and unwitting bias are two important difficulties in the process.

Observation may yield classificatory and descriptive symbols as well as rank symbols but seldom may it produce scale numbers. Often, the immediate result of observation is an evaluation, not a measurement. Several paper-and-pencil devices are available for the recording of observations, for control of the observer's attention, and for protection against bias. Among these are check lists of dimensions to be observed, anecdotal record forms, and various types of rating scales.

The validity and reliability of observations are increased by the following observances. Use observation as a measuring procedure where and when legitimate only. Use appropriate paper-and-pencil devices and record quickly. Guard against bias and whenever possible base evaluations on several observations.

EXERCISES

1. Along with several other students, observe a group of children at play or at school work and make an anecdotal record of what you see and hear for fifteen minutes. Compare your record with the records of your classmates. Note all the differences among your records and discuss the reason for them as well as their implications for reliability in observation.
2. Select some phase of a subject you teach or intend to teach for which pupil evaluation might be based on observation. Prepare a check list, an anecdotal form, or a rating scale for use with pupils on this phase of the subject.
3. Restricting yourself to a grade or to a subject, list the aspects of pupil performance and achievement best measured or only measurable through observation. Briefly justify the use of observation for each.
4. Reflect on and list the biases you have that might affect your evaluations of pupil conduct or citizenship. Indicate how you might guard against the influence of each of the biases you list.
5. Inspect the guidance folders of several elementary and secondary pupils. Notice all the observational records and ratings contained in them. Write a brief critique of the observational procedures, based on the principles of valid observation stated in this chapter.

BIBLIOGRAPHY

1. Buros, Oscar K., *Fourth Mental Measurements Year Book*. Highland Park, N. J.: Gryphon Press, 1953.
2. "Evaluating Pupil Progress," *California State Department of Education Bulletin*, XXI, No. 6, April, 1952.
3. Haggerty, M. E., Olson, W. C., and Wickham, E. K., *Haggerty-Olson-Wickham Behavior Rating Schedules*. Yonkers, World Book Co., 1930.
4. Keisler, E. R., "An Improved Formula for Scoring Certain Guess-Who Ratings at the Adolescent Level," *Journal of Educational Psychology*, 45:151-160, March, 1954.
5. Klein, Arman, and Thomas, Leah C., *Posture Exercises*, Children's Bureau Publication No. 165. Washington, D.C., U.S. Department of Health, Education & Welfare, 1926.
6. Seedorf, E. H., "Experimental Study in the Amount of Agreement among Judges in Evaluating Oral Interpretation," *Journal of Educational Research*, 43:10-21, 1949.
7. Sells, S. B., "Observational Methods of Research: rating scales," *Review of Educational Research*, 18:429, December, 1948.
8. Wilson, J. W., "Correlation of Clinical Estimates with Test Scores on Mental Ability and Personality Tests," *Journal of Clinical Psychology*, 10:97-99, January, 1954.
9. Wrightstone, J. W., "Rating Methods," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Co., 1950, pp. 961-964.

CHAPTER 5

PRODUCT ANALYSIS AND FREE-RESPONSE PROCEDURES

A great amount of educational measurement is based on appraisals of products. Pupils write compositions, make drawings, prepare notebooks, and design various artifacts in such classes as shop and homemaking, all as part of their regular learning activity. Then teachers examine these products to evaluate the pupil's achievement. Again, pupils are directed to produce things specifically for purposes of evaluation: to write paragraphs or short compositions on specified topics, to draw a paramecium as part of a biology test, or to sketch a cross section of a typical volcano in a geography quiz. Such "test products" constitute an additional important aspect of educational measurement.

For semantic convenience "test products" are called free responses in this text. The more usual term for a test item that asks for an extended written response is "essay question" or "essay examination." This term has unfortunate connotations of unreliability and, moreover, is too restrictive for our scheme of classification. Consequently, the phrase "free-response" question or item is preferred. As will be seen in Chapter 6, free response is to be contrasted in our view with *guided response*, a generic heading for true-false, multiple-choice items and the like.

In this chapter we shall treat product analysis and free-response procedures jointly because of their inherent similarity as measuring procedures. In turn, attention will be given to some of their general characteristics, to types of free-response questions, to methods of scoring and forms of measurement and, finally, to the applicability of the procedures.

General Characteristics of Product Analysis and Free-Response Techniques

In Figure 6 are portrayed some representative products together with scoring notes, and in Figure 7, examples are shown of free-response items and their scoring. Both sets of examples have been selected to illustrate the varied approaches to and problems inherent in such measuring procedures. Frequent reference will be made to the Figures as our discussion progresses.

Products

Scoring

Over-all ranking

In assigning this a rank the teacher simply compared the drawings of all his pupils, placed them in order of merit, and gave the drawing the appropriate number. The ranking is entirely subjective but is based on his opinion of such factors as composition, drawing skill, perspective, and sympathy with the subject.



A.

Over-all rating

The teacher of this pupil in an eleventh-grade U.S. history class rated his cartoon as an *A*. This means that the teacher had in mind several categories of cartoons and the pupil's had the essential characteristics of the *A* category. These characteristics are a clear conception of a true and important historical situation that is illustrated clearly, appropriately, and interestingly.



B.

Figure 6. Examples of pupil products scored by various methods.

Products	Scoring
C <i>Spelling errors 3</i> <i>Punctuation " 3</i> <i>Usage " 3</i> <i>The Air Force Flying Saucers.</i> <i>I in the last eight years 6,000 saucers sightings have been reported to the Air Force. But the Air Force hasn't found any evidence that seems to indicate any saucers were being. As this statement was made by the Air Force also gave evidence of supporting a plan for building a flying saucer. This saucer will be built by Aero Aircraft, Ltd., near Toronto, at a cost of 100 million dollars. I actually saw one in a picture photograph of the saucer and although the Air Force released an official statement of the saucer. It is supposed that it will be a one man fighter in the 1,000 miles per hour class. It will take off vertically, as the jet exhaust is spread around the outer edge of the saucer and to give it the vertical position.</i>	<i>Factor counting</i> The teacher was concerned with three types of error: in spelling, in punctuation, and in syntax. The errors in the pupil's paper were classified and counted. Their kind and amount made the paper barely satisfactory, so the pupil was given an S—.
D 	<i>Factor rating</i> This dress, made by a Grade XII student, was rated on a five-letter scale (A, B, C, D, F) for each of four factors, as follows: 1. Suitability of pattern and material B 2. Construction (neatness, accepted methods, pressing, time to complete, etc.) C— 3. Appearance on individual C 4. Product as compared with girl's ability C—

Figure 6. (Continued)

Free-response items	Scoring
Question: What is the difference between an electromagnet and a permanent magnet? What are some uses of electromagnets? Answer: An electromagnet is an iron 2 core with a wire around it and an electric current goes through the wire. 2 The difference is an electromagnet can be turned on and off by electricity. 2 Some uses are a doorbell, crane, 3 motor. <u>9</u> E.	<i>Factor counting with weights</i> Each specific and accurate characteristic of electromagnets against permanent magnets was to receive 2 points. Each accurate use was to receive 1 point up to a maximum of 5. The different weighting for characteristics and uses indicates that knowledge of characteristics is thought to be the more important.
Question: What is the length and width of a rectangle whose length is 2 inches more than its width and whose perimeter is 40 inches? Answer: - Given: Length is 2 inches greater than width. - To find: Let width = x Let length = $x + 2$ inches - Conditions: - Equations: $2x + 2(x + 2) = 40$ $2x + 2x + 4 = 40$ $4x + 4 = 40$ $4x = 36$ $x = 9$ $x + 2 = 11$ - Solution: Width = 9 inches Length = 11 inches F.	<i>Factor rating (as right or wrong)</i> The purpose of this algebra problem is to see whether pupils can use a given system for analyzing and solving a problem. There are five steps in the system and the pupil is marked right or wrong for each step. This pupil was in error on Step 1, omitting the perimeter of the rectangle, and he omitted Step 3 entirely.

Figure 7. Examples of free-response items scored by various methods.

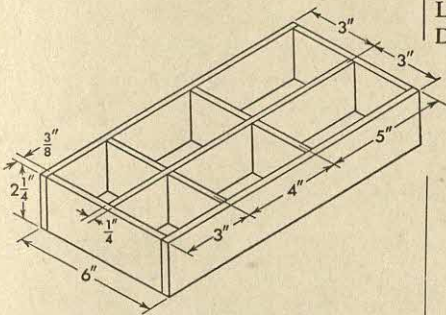
Free-response items	Scoring
Directions: Make a working drawing of a nail box. 	<i>Factor rating</i> The pupil was rated according to a five-letter scale on different aspects of his work, as follows: Lettering <i>B</i> Neatness <i>A—</i> Accuracy <i>A</i> Line work <i>A</i> Dimensions <i>B—</i>
Directions: Write the first two sentences of the Gettysburg Address in your very best hand. Use pen and ink. <i>Four score and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war testing</i>	<i>Comparison with a product scale</i> The pupil's specimen was compared with the several samples on the Ayres handwriting scale. His writing most closely resembled that in the sample rated as 60. Hence his handwriting was scored as 60.

Figure 7. (Continued)

By their nature, neither products nor free-response items involve standardized responses. In both instances the variety of possible responses is virtually unlimited and automatic scoring thus is precluded. Validity and reliability are to be gained through proper use of standard analysis, sampling and standardized stimulations.

This matter of standardized stimulations is the only important difference between product analysis and free-response techniques as measuring procedures. Test questions and the test situation tend to constitute more controlled and uniform stimulations for all students than do assignments and study situations. As a corollary, answers to "essay questions" tend to be shorter and more stereotyped than compositions on the same subject written for instructional purposes. Hence, free-response techniques, as against product analysis, are likely to produce the more comparable measures.

Substantial and important similarities exist between both these procedures of measurement and those of observation. Products and free responses are appraised directly with unaided senses and often they are evaluated quickly and subjectively just as are the behaviors measured through observation. Please notice that for all the examples in Figures 6 and 7 the scorer must read or view the products himself, and that for Examples *A* and *B* the scorer enters entirely subjective ratings for the products as a whole. In product analysis and free-response techniques, the things measured are artifacts. They do stand still and, thus, can be reappraised. This amenability to reappraisal makes the two techniques potentially more reliable than observation, but bias and other irrational factors still can influence the measures obtained from them, even as they distort the results of observation.

Tenets of Good Observation Pertinent to Scoring Products and Free Responses. To insure that direct and subjective evaluation does not destroy the effectiveness of product analysis or free-response techniques, it is necessary to follow the principles laid down for observation (pages 60-61). In brief, they are:

1. To keep measurement and evaluation separate
2. To use a known and defined criterion
3. If possible, to use a tangible criterion. Such a criterion is illustrated in Example *H*.

The effect of personal bias is to be avoided, either through self-control or through compensation. In Example *D*, for instance, the homemaking teacher may have had a preconception that plain but expertly sewed dresses are better than fancy ones. She may, by concentration, attempt to hold this bias in check when she appraises the dress or she may use a system of scoring that compensates for its operation, perhaps weighting other factors more heavily. The irrational influence of fatigue, wandering attention, shifting point of view, etc., similarly may be minimized by systematic scoring and by concentration on the task.

Products as Evidence or Having Self-Significance. Products and free responses may themselves be the phenomena whose measurement is desired (notice Examples *A*, *D*, and *H*), or they may be construed as evidence of some ability or knowledge (Examples *B*, *C*, *E*, *F*, and *G*). The painting, the dress, and the handwriting specimen have significance in themselves, and further inferences based on them about pupils' ability or skill may not be necessary. On the other hand, cartoons about history, solutions to mathematics problems, and answers to science questions usually have no significance in themselves. They serve only to indicate knowledge of history, ability in mathematics, or comprehension of science. When the products and free responses are themselves the objects of measurement, maximum validity is to be expected. But when they are measured merely as evidence of some pupil attribute, their validity is lessened. This is true because in addition to errors in defining, detecting, and measuring dimensions of the product there are likely to be errors in relating the dimensions of the products to the dimensions of knowledge, skill, etc., of which they constitute evidence.

Eliciting Free Responses

Free-response tests are utilized because it is anticipated that free responses to test items more nearly approximate a pupil's natural actions than do his responses to objective or guided response tests. Thus, a first step in designing a free-response procedure is to decide upon a "natural" product for which measurement is appropriate, either because the product itself is educationally important or because it is demonstrative of something else (a knowledge, attitude, etc.) that is important. A question or direction for free response is to be designed, then, which will draw forth all or a portion of the selected "natural" product or something that will approximate it.

To understand this relationship between free response and natural product, please notice Example *F* in Figure 7. Why, we may ask, is it significant to give a pupil an arithmetic problem about the dimensions of a parallelogram? The rationale behind use of a thought problem as a testing device in mathematics may run something like this. Pupils encounter mathematical situations in their everyday life and respond to them with thought-out solutions or with written-out solutions. The conglomerate of any pupil's solutions is what we mean by his "knowledge" of mathematics or his mathematical "ability," and the average competency of his solutions is the degree of his knowledge or ability. Moreover, the quality of any given solution is considered to be similar to the quality of all the others. By giving a pupil a mathematical problem on a test, we can derive from him a "test product" that will be like one of the "natural products" collectively comprising his mathematical ability.

To produce free responses that approximate natural products (in essential respects, of course) a number of verbal and graphic devices are available. These are the standard stimulations of free-response techniques. In addition to their basic requirement for stimulating the desired type of response, it is essen-

tial that they constitute the *same* stimulation for *all* pupils who react to them. This means that they must be simply and clearly phrased; that they must avoid ambiguous words and constructions; and that they not assume special experience on the part of pupils if they are addressed to an unselected group of pupils.

Directions to Write, Draw, Design, or Make Something. These are exemplified by Examples *G* and *H* in Figure 7, and, along with questions, are the most usual type of free-response item used for measuring educational achievement.

Questions. Examples *E* and *F* in the same Figure 7 illustrate questions. Such interrogatory free-response elicitors are the "essay" questions of long educational use condemned by some advocates of "objective" measurement for their "subjectivity."

*Open-End Statements*¹. In the attempted measurement of personality factors (interests, fears, etc.) it often is desirable that a person be permitted to write whatever he first associates with a given stimulus. To elicit such free associations, open-end statements and stimulus words, objects, and pictures are used extensively.

Examples of open-end statements are:

My school is _____

I hate to _____

The mother and father _____

*Stimulus Words and Objects*¹. These often are used for verbal response in counseling interviews, in psychiatric examinations, and are an integral part of lie detection procedures. Examples are:

Write down (or say) the *first* idea that occurs to you when you see (or hear) each of the following words:

Butter _____

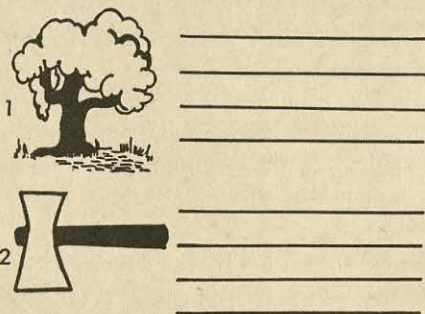
Medicine _____

Woman _____

Crime _____

¹ NOTE. The use of open-end statements and stimulus words, objects, and pictures in personality measurement requires extensive special training. They should not be devised, nor responses to them interpreted, by persons who lack this training, however skillful they may be in measuring educational achievement. As often as not, the responses elicited are verbal rather than written. Projective techniques are discussed at greater length in Chapter 15.

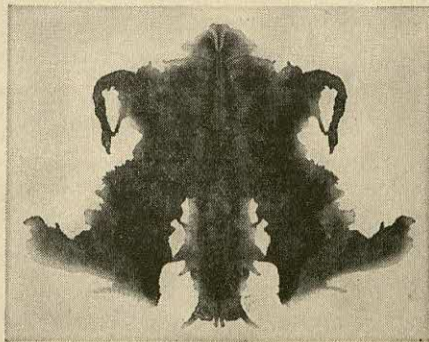
Write whatever comes *first* into your mind when you see each of these objects.



*Stimulus Pictures.*¹ Pictures are another type of stimulus for free association used in personality measurement. The best known are the Murray Pictures, examples of meaningful stimuli, and the Rorschach Ink Blots, examples of abstract or nonmeaningful stimuli. A sample of each is shown in Figure 8.



A Murray Picture



A Rorschach Ink Blot

Figure 8. Sample pictures from the Murray Thematic Apperception Test and the Rorschach Ink Blot Test (5). (The former is reprinted by permission of the publishers from Henry A. Murray, *Thematic Apperception Test*, Cambridge, Mass., Harvard University Press, Copyright 1943 by the President and Fellows of Harvard University. Plate IV of the Rorschach Test, by courtesy of Hans-Huber, Publisher, Berne, Switzerland.)

In addition to their use in personality measurement, pictures as stimuli to free responses frequently are used in intelligence tests. As a rule, such pictures suggest a given story or event, and a pupil's response is judged to be adequate as he approximates the proper story or event.

Stories to Complete. A little used but potentially valuable free-response elicitor is the incomplete story. This is found in some intelligence tests and most frequently in Language Arts instruction. For example, to gauge pupils' skill at narration, a teacher in a tenth-grade English class writes on the blackboard:

"William Hunt planned to go on a bicycle hike with his gang one Saturday. His older sister teased him when he went to bed very early Friday night so that he would have lots of energy on Saturday. During the night he heard his dog barking once in the back yard. When he awoke on Saturday, he dressed for the hike and even before breakfast prepared himself several sandwiches for a roadside lunch. He took these together with a poncho (in case of rain) out to his bike to tie on the luggage carrier. *But his bike was gone!*"

"(Each of you finish this story just as you think it ought to be finished.)"

Problem Situations. The problem situation has been used more and more extensively in recent years by Social Studies teachers to test pupils' ability to think critically. Mathematics teachers have been using the device for many years whenever they have given a thought problem. In substance, a problem situation is a statement of variables and the need for their proper resolution. The individual is required to examine and relate the variables in an appropriate way and present the right or a right solution. Example *F* in Figure 7 is one example of a problem situation. Another is this item from a civics test appropriate to the eleventh or twelfth grade.

"Some cereal companies claim that their cereals contain 'anonoxin,' which prevents tooth decay. Some dentists are reported to have said that 'anonoxin' does prevent tooth decay. It is known, of course, that sugar and soft foods are factors in tooth decay, and that cereals are largely starch which becomes sugar in the mouth, and that they need little chewing. You are a teacher and a representative of one of these cereal companies wants you to use a movie produced by his company. The film is on health and, among other things, it says that 'anonoxin' prevents tooth decay. It does not advertise any cereal but merely says at the end that ——— cereal company made the picture."

"What should you do and why?"

Requirements for Free-Response Elicitors. Any of these free-response elicitors—questions, directions, stimulus words, etc.—must not in itself indicate what response is expected. Differences in knowledge or intelligence or attitude among pupils should be what makes for their differential responses, not their ability to read. On the other hand, free-response elicitors should indicate clearly the type of response desired and its limits as to time, space, or detail.

Without these indications, pupil responses are not going to be comparable. To illustrate the desirable middle ground between items that contain clues to the right answer and those entirely too vague or unlimited, consider the following:

Too many clues Describe the development of the "Bill of Rights," stating its first origins in England, its relation to other documents, and how it became "attached" to the Constitution.

Too vague Discuss the Bill of Rights.

About right Describe the important events that made the "Bill of Rights" part of our basic national law.

In preparing tests to contain free-response items, it is essential to leave appropriate space for the responses. Pupils tend to gauge the extent of their responses by the space afforded them. If a few phrases are all that are expected in response, a half page of space is likely to make conscientious pupils strive for additional statements and perhaps instigate repetition, padding, or even error. On the other hand, if many sentences are needed for an adequate response, it is frustrating to the pupil to have only two lines in which to reply. In his frustration the pupil may abbreviate, cramp his writing, omit factors, or doubt his interpretation of the question, all to the detriment of his response.

Scoring Products and Free Responses

The assignment of proper measurement symbols (numbers or letters) to products and free responses is, of course, the critical phase in these measuring procedures. The symbols most frequently applied are classificatory in nature or are indicative of rank. Scale measurement usually is not possible except in research situations where special devices are used. In the example items shown, the painting in Example A has been given directly a rank symbol while the cartoon in Example B, the dress in Example D, and the drawing in Example G have been assigned a letter indicating their classification or rating. In Example H, the handwriting specimen has been assigned a classification number that may be interpreted as a rank number. The numbers found for Examples C, E, and F are still raw scores and may be converted into either class or rank symbols.

With scores or marks restricted largely to the two least precise forms of measurement, careful scoring of products and free responses is essential so that further imprecision is held to a minimum. The different scoring methods

in general use include many that have such necessary precision, but, unfortunately, some also are widely employed which lack it. As a rule, scoring products or free responses as a whole is a less reliable approach than scoring them piecemeal or factorially. The purpose of measurement and the time available for it may permit or require use of over-all scoring, but, if they do not, some sort of factorial scoring is thought to be essential.

OVER-ALL RATING

The scoring of the cartoon in Example *B*, Figure 6, illustrates the direct assignment of a classification symbol to a product or a free response. The scorer in this case simply viewed the cartoon, decided subjectively that it belonged to a given class of cartoon, and assigned the letter *A*, which symbolized the classification. It might have been any other letter or number, or even a word, that denoted the classification. The important thing is that the cartoon was considered to have characteristics similar to the classification.

Of all methods of scoring products and free responses, this is inherently the most unreliable, albeit the most rapid. A good deal of its unreliability may be allayed if the scorer has adequate directions, and if he follows the directions in the same way for each product or free response he marks. The directions (prepared by the scorer himself as a rule) should specify the aspects of the product significant for measurement. They should represent the classifications to be applied, and should contain examples of products that typify the several classifications. (The Binet manual contains excellent examples of directions for over-all rating of free responses [7].) Fatigue, boredom, and deadlines are among the principal reasons for uneven reaction to a series of products or free responses. Consequently, scoring should be attempted only when, and for as long as, one can sustain interest in the task and remain untired. To avoid having to mark the last of a series in a rush, the apparent solution is to allot sufficient time in advance or, failing this, simply to miss the deadline. Principles stated for the use of behavioral rating scales (page 59) are applicable to the rating of products.

OVER-ALL RANKING

In Example *A*, the pupil's painting has been measured by comparing it with the water colors of other pupils, and then by assigning it an appropriate rank in this group of paintings. In art contests evaluations frequently are made simply as a result of ranking the art work submitted.

Where products or free responses are to be compared in one dimension only, the method of over-all ranking can be highly reliable, far more so than any over-all rating. The reason for this is obvious from the experience of any of us. Consider, for example, the judgments we make about musical tones. We can "rate" tones heard by calling them *a*, *b*, *c*, *d*, etc., a classification form of measurement. But only trained musicians can do this with any degree of accuracy. Or, we can say simply that this tone is higher, this is lower, and by continuing our comparisons, establish a rank order of tones from lowest to

highest. This any of us but the tone-deaf can do and with greater accuracy than we can classify the tones *a*, *b*, *c*, etc.

When several dimensions are to be the basis for comparison, reliability is lessened and the method, at some undetermined number of dimensions, becomes as unreliable as over-all rating. This may be seen by considering thirty short compositions written by seventh-grade pupils. If these are to be ranked simply on the basis of proper usage, ranking is easy and fairly precise. But if they are to be ranked on usage plus penmanship, a hypothetical average of some sort must be struck between the two, for each paper and the papers compared in relation to this average. If to usage and penmanship is added interest, even more vague averages must be estimated for the paper; and if a fourth dimension is to be reckoned with, the teacher may with good reason decide to give up trying to rank them.

To obtain maximum reliability in over-all ranking, it is necessary then to assign rank for one dimension only, and, obviously, it is necessary that this dimension be clearly defined. In addition, certain other procedures have been found to add to the reliability of the method. First, all the products or free responses to be ranked should be read or viewed *before* any can be ranked. If there are many products in the group (say 15 or more, as a rule of thumb) it is advisable to compare each with every other, or at least to make such item-by-item comparisons among any subgroup greatly alike. Even more reliability may be attained by a second determination of rank and a reconciliation of differences between the first and second runs. In research, and certainly as the basis for critical decisions, ranks assigned independently by several impartial but equally qualified judges should be pooled to determine final rank.

It may be apparent by this time that over-all ranking (though inherently a more reliable method of scoring) is likely to be far more time-consuming than over-all rating. If done as quickly as over-all rating often is done, there is no certitude that it will be any more reliable.

COMPARISON WITH A PRODUCT SCALE

We have just established that product-by-product comparisons may be made with some accuracy, but we also have observed that the process is time-consuming, and that it yields rank symbols only. A method of comparison more rapid and capable of yielding classification symbols as well as rank numbers should then be of great value. Such a method is the *product scale*.

The handwriting scales in use in elementary grades are the foremost instances of this method of measurement (notice Example *H* in Figure 7). Stereotyped handwritten passages have been developed that represent the variety in pupil handwriting in a number of gradations, from the least legible and attractive to the most. A pupil is asked to write the passage used in the scale in his own hand. The teacher simply finds on the scale the specimen most like the pupil's, and the pupil's paper is given the number or letter of this specimen.

Except in handwriting, there are few published product scales. It is difficult

to standardize them, and in American schools most subjects lack the standardization of content that might make them useful. The virtue of a published scale is that its numbers may indicate nationally significant classifications or rank in a norm group. In some scales an effort has been made to have specimens typical of various grade or age levels. In others, equal differences in skill or difficulty have purportedly been established among all the specimens contained in the scale.

For subjects in which pupil products are likely to be of the same type for a number of years, it is possible for a teacher to develop his own product scales. To do this, you need to select from each group of products submitted by pupils several that are representative of the worst, best, and intermediate levels of attainment. When selections have been made from three or four such batches, establish a rank order among all the selections, using fellow teachers as judges in addition to yourself. Reject all specimens upon which there is disagreement as to order. The remainder is your product scale. You may assign numbers or letters to the scale specimens and use your scale as you would a published scale. *Do not, of course, attribute scale significance to numbers assigned to specimens.* If the scale seems to have too many units, select 5, or 7, or 10, or whatever number seems best but be sure that these products are evenly spaced along the larger scale.

FACTOR RATING

If a product or free response has more than one measurable dimension (as they generally do), and if all the dimensions are to be reckoned with in scoring the product or free response, both over-all rating and over-all ranking have serious limitations. As we have observed, you are forced to derive mentally some average among the dimensions, and actually to rate or rank this average (which is only an idea) rather than anything in the product itself. Because this process is subjective and unsystematic, an indeterminate degree of error is certain. To avoid this error, it is advisable to use some method of analytic scoring.

This method is exemplified in Examples *C, D, E, F, and G*, in Figures 6 and 7. Notice that in each case several things are measured, not just the thing as a whole. Factorial scoring consists of the identification of factors that constitute the important properties of the product, the separate measurement of each such dimension and, if necessary, the systematic combining of the separate measures into a single one. In factor rating (the method of analytic scoring to be discussed first) the dimensions are immediately assigned a measurement symbol, usually a number, which represents a classification or a point on some imaginary scale.

Advantage of Factorial Scoring. The advantage of dealing with dimensions separately is that each measure then represents the status of a given dimension and that only. In over-all scoring, on the other hand, the same measure may represent different degrees of several dimensions and be indica-

tive only of their hypothetical average. To see this distinction, consider the case of English compositions, two of which may have been given directly an over-all score of 50. The fifty for one paper could represent a hypothetical 70 in grammar, 30 in style, and 50 in content, whereas the 50 for the other could stand for a hypothetical 50 in grammar, a 70 in style, and a 30 in content. With factorial scoring, the different numbers would first have been actually assigned to the three dimensions and the different status of the different dimensions of the compositions then characterized as different. If a single number were needed to represent each composition, the separate measures of the three dimensions would still remain on the papers to explain what the 50 really meant.

It must be recognized, of course, that proper scoring of a product may require attention to its over-all pattern or Gestalt as well as to its separable dimensions. If so, this over-all aspect should be defined as a separate dimension and measured as any elemental dimension. Moreover, just because there may be such a holistic dimension is no reason to disregard factorial scoring.

Once the dimensions have been identified, it remains to rate them in appropriate ways. The general principles and rules for the use of behavior rating scales (page 59) are as germane to factor rating as they were to over-all rating of products. Numbers or letters assigned to dimensions as their ratings are to be viewed as classification symbols only. If a single rating is to be derived from the dimension ratings, numbers must be assigned to the dimensions rather than letters or words. The single score may be a total or an average.² While precedent and administrative convenience may require such single ratings, they obscure as much as they reveal and *you are advised not to use them for instructional purposes*. The factor scores themselves tell the pupil his strengths and weaknesses and show the teacher in detail what he has accomplished and what he has not.

Example *F* in Figure 7 illustrates a special case of factor rating, the "correct" or "incorrect" assigned to aspects of solutions to mathematics problems. In this case, it is assumed that only one type of response is correct for each phase and all others are incorrect. It remains for the scorer to look only for one class of responses, correct ones. He need not discriminate among the others. The distinction between a correct response and an incorrect one presumably is clear cut and constant, and if this assumption is borne out, this special type of factor rating can be highly reliable.

FACTOR COUNTING

A more reliable method of factorial scoring capitalizes on the precision of enumeration. You know, of course, that we make fewer errors in counting than we do in most of our perceptual activities. In fact, counting is about the

² It is considered mathematically invalid to apply addition and division to classification numbers (see page 9). However, custom has long condoned and probably will continue to condone the practice in the scoring of products or free-response test questions.

only thing man is still trusted to do in the physical sciences. His color discrimination has been replaced by spectography; his sense of temperature, by the thermometer; his auditory discrimination, by electronic devices; and, of course, his senses of distance and weight by calipers, meter sticks, and scales. But he still enumerates many things in the laboratory and his counts are admitted as scientific data (if he is properly trained, of course).

Advantage of Factor Counting. In addition to relative freedom from perceptual errors, enumeration usually can avoid the influence of bias and other subjective sources of error. It is relatively easy to say that this is a "good" paper to a pupil whom you like, that this is a "bad" paper to a pupil whom you dislike, and yet have the papers as judged by others be equivalent in value. It is relatively difficult on the other hand, for your prejudices to affect your count of spelling errors in the two papers. Moreover, counting may be continued with little loss in accuracy despite fatigue, boredom, or malaise, while these same things can make qualitative appraisals entirely unreliable. In fact, the stigma of subjectivity usually is lifted from behavioral measurement when simple enumeration is involved.

Dimensions Must Have a Quantitative Aspect. Factor counting requires first that each dimension to be measured be defined in terms of things that can be counted. In Example C in Figure 6, good grammar was defined as the absence of three types of error: punctuation, spelling, and syntax, all of which errors can be noted and counted. Moreover, unless a dimension *can* be defined in terms of things subject to enumeration, it must be rejected if factor counting is to be the only scoring method applied.

For each product, we count the factors that belong to each dimension of concern. These numbers may be left as such, as raw measures of the dimensions, or they may be converted into rank or other derived numbers.³ If necessary, they may be combined into a single score for the product just as were factor ratings.

Applicability of Factor Counting. A great many of the educationally significant dimensions of products and free responses are amenable to factor counting. Some of these, along with illustrative items to be enumerated, are shown in Table 1. Other dimensions (not so readily susceptible to a counting approach) may, with ingenuity, be so defined. For hundreds of years literary critics have insisted that a writer's style can be described only in qualitative and figurative terms. Now, authors of formulas for judging the difficulty and interest of books are asserting that many enumerative items collectively constitute style, i.e., length of words, length of sentences, frequency of personal referents, etc.

Factor counting, as a method of scoring products and free responses, has such a great potential for increasing the reliability of these procedures that it deserves the widest possible use. Both product and free response appraisal fell

³ The conversion of raw scores into rank indexes or other derived numbers is discussed on pages 155-160.

TABLE 1
Some Pupil Products Particularly Amenable to
Factor Count Scoring

<i>Products</i>	<i>Illustrative factors to be counted</i>
English compositions or themes	Usage { - Spelling errors - Punctuation errors - Word form errors - Sentence errors
	Style { - Different sentence forms - Mean sentence length - Parallel constructions - Figures of speech - Clichés - Repetitive adjectives
Reports and answers to free-response questions in:	Facts or accurate ideas Correct statements of relationship Correct generalizations
Civics History Geography General Science Biology	

into disrepute as means of measuring educational achievement because of their unreliability. Their intrinsic validity has seldom been questioned. Consequently, if their reliability can be increased to approximate that of the "objective" test, an excellent educational tool has been reclaimed.

FACTOR WEIGHTING

A frequent criticism of factorial scoring is that trivia may be given undue significance and critical points may be slighted, because all are added together to produce the single number that is the product's or free response's score. This condition may be avoided, first, by not deriving a total score. But if one is needed, the relative significance of different factors may be retained by some system of factor weighting.

Importance Weighting. Two methods of weighting are in common use. One bases the weight of factors on their estimated importance, and this method is the one illustrated in Example *E* in Figure 7. In this example, as in all cases of importance weighting, the different numerical values assigned are arbitrary and the decision that one dimension is more important than another has little empirical justification. If, of course, many teachers and authors in the field have agreed on a relative order of importance, there is at least the authority

of consensus on which to depend. Again, should certain dimensions be composites of a known number of lesser dimensions, there is some logical justification for weighting the composite dimension with a number equal to its components. However arbitrary and subjective importance weighting may be, if done carefully, and frequently corrected, it helps to make single scores for multidimensional items more valid.

Cook has derived from many sources certain criteria for the significance of items of learning that may have a bearing on factor weighting. These are "(a) the frequency with which it will be used, (b) the cruciality of the situations in which it is needed, (c) the extent to which superior individuals use it, (d) the universality of the need in different vocations, (e) the universality of the need in different geographic areas, (f) the universality of the need in different time periods, (g) the difficulty of learning, (h) the frequency of error, and (i) the frequency of use in a specific vocation" (1).

Difficulty Weighting. The other method of weighting is a function of the difficulty of the dimension. Difficulty being applicable only to tests of skill, knowledge, or intelligence, this method is usable only in these areas. The relative difficulty of factors may be based on opinion, as it is with importance, or it may be based on the performance of pupils with respect to the items. The latter procedure is the more objective and systematic of the two and, hence, the preferable one. It is explained by means of the following example.

In a biology class, the teacher recurrently tests pupils by asking them to describe an animal typical of the life form they are studying at the moment. Over a period of years he has classified and tabulated the statements they include in their answers and determined for each classification the percentage of statements that are correct. His cumulative tabulation resembles this.⁴

	<i>Per cent of pupils who answer correctly</i>
Classification	85
Essential physical characteristics	70
Significance in human affairs	50
Method of reproduction	45
Type of food, hosts, or prey	35

On the basis of these percentages, the teacher gives different weights to correct statements according to the factor they represent. The weights are:

Classification	1
Physical characteristics	1

⁴ The percentages are illustrative only, and do not indicate the actual relative difficulty of these factors.

Significance in human affairs	2
Method of reproduction	2
Type of food, host, or prey	3

Any pupil's total score on the test then becomes the sum of the correct statements multiplied by their proper weights. A pupil's paper might produce a score of 17 by stating 3 correct items of classification, 5 of physical characteristics, 2 about the animal's social significance, 1 on reproduction, and 1 on food, thus:

$$3 \times 1 = 3$$

$$5 \times 1 = 5$$

$$2 \times 2 = 4$$

$$1 \times 2 = 2$$

$$1 \times 3 = 3$$

$$17$$

General Significance of Weighting. The principle of weighting is as applicable to a test as a whole as it is to the answer to any free-response item. Time or length of response as well as importance and difficulty may be used to determine proper interitem weight.

ADDITIONAL PRINCIPLES FOR SCORING PRODUCTS AND FREE RESPONSES

From the descriptions of over-all rating and ranking and of factor rating and ranking as methods of scoring, you may have inferred that each method is to be used separately and exclusively. To the contrary, effective scoring may require a combination of methods since products and free-response tests are not apt to fit any given stereotype of scoring. Hence, it is advisable to use a scoring system that is best adapted to your purpose in measurement and to the types of products or free responses you intend to appraise. Combine factor counting and factor rating procedures as you need to handle different types of dimension. If some free responses are unidimensional, while others are multidimensional, use over-all rating as well as factor rating. The ultimate criterion for product and free-response scoring is that measurement symbols be assigned to them that characterize their status most precisely, not that a given method of scoring be used.

The system of scoring devised for a given free-response test or product is the scoring "key." It will not permit automatic scoring as will the key to a guided response test, but it will insure maximum reliability in ratings, rankings, and counts. The key should contain as a minimum the numbers or other symbols to be assigned, the dimensions to which they are to be assigned, dimension or item weights, the significance of varying numbers or letters, and a formula for a total score, if one is to be derived. Where products or free responses relate to an area of knowledge, the elements of knowledge repre-

sented constitute the most important dimension of the products or free response. Consequently, the facts or relationships that would comprise an ideal product or response should be written down. They will serve as a reference point for rating answers or as reminders of correct factors if a counting procedure is used.

In scoring free-response tests, it is advisable to score all papers on one question, then all papers on the second one and so on, rather than score one pupil's paper in entirety before scoring the second paper. By this means, attention may be kept on a single focus and greater speed is possible. Moreover, each pupil's response to any question is more likely to receive equitable scoring.

Applicability of Product Analysis and Free-Response Procedures

Any phenomenon of educational significance written, drawn, or made by a pupil, or closely related to something so written, drawn, or made, may be measured through product analysis and free-response tests. The procedures then are widely applicable and are used in all school grades and subjects.

Product analysis is the primary means of measurement in shop instruction, in homemaking and art, and is of great importance in language arts and social studies. Free-response procedures are employed extensively in social studies, language arts, mathematics, and science. In addition, both products and free responses are used by home room teachers, counselors, and school psychologists to appraise personality variables. Pupil "autobiographies" and paintings are the two products most widely used for this purpose.

The two procedures are recommended where they are applicable because of their inherently high validity. As we have seen, their reputed unreliability is as much a function of the way they are used as of the techniques per se. Properly used, they can be highly reliable. A free-response procedure should be selected in preference to product analysis if strictly comparable measures are necessary for a group of pupils.

Summary

What pupils produce in school—compositions, drawings, notebooks, tie racks, and dresses—often are employed in educational measurement as well as are the free responses they make to certain test questions. Such instructional or test products are appraised directly with unaided senses and often are evaluated quickly and subjectively. Thus, product analysis and free-response tests are obedient to much the same principles as observation.

Free responses in tests are elicited by directions to do something, by questions, by open-end statements, and by stimulus words and objects, stimulus pictures, stories to complete, and problem situations. Both products and free responses are susceptible to various methods of scoring. In reverse order of reliability these methods are over-all rating, over-all ranking, comparison with a product scale, factor rating, and factor counting. The last two methods in-

volve the identification of the dimensions that comprise any product or test response and the separate rating of each or the separate enumeration of the components of each. In factorial scoring, weights of different size may need to be assigned to dimensions or the elements thereof that are of varying importance or difficulty.

In practice, the scoring of products and free responses may require a combination of methods. The system of scoring devised for any product or free response item is its "key" and should be rigorously followed in scoring. In scoring free-response tests, it is advisable to mark all papers on one item, then all papers on a second item, and so on, rather than to score all of one paper before scoring the second paper.

EXERCISES

1. List the pupil products that are of importance to the subject or grade in which you specialize, and state for each several dimensions that should be evaluated.
2. Prepare a free-response test of at least five items for the subject or grade in which you specialize. Devise a "key" for scoring responses to each question.
3. Obtain sample pupil products for the grade or subject in which you specialize and score them first by over-all rating and second, by factor rating or counting.
4. Put these same products in order of excellence from best to worst. Select at least five that could constitute a rough product scale.

BIBLIOGRAPHY

1. Cook, Walter, "Tests, Achievement," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Co., 1950, p. 1466.
2. Cronbach, Lee J., *Essentials of Psychological Testing*. New York: Harper & Bros., 1949, pp. 294-298.
3. Murray, Henry A., *Thematic Apperception Test*. Cambridge: Harvard Press, 1944.
4. National Society for the Study of Education. "The Measurement of Understanding," Part I, *Forty-Fifth Yearbook*, 1946.
5. Rorschach, Herman, *Psychodiagnostic Plates, Fourth Edition*, 1921-1945, and *Psychodiagnostics, Fourth Edition*, 1949, New York: Grune and Stratton, Inc.
6. Sims, V. M., "Essay Examination is a Projective Technique," *Education Digest*, 14:38-31, 1948.
7. Terman, L. M., and Merrill, Maude, *Measuring Intelligence*. Boston: Houghton Mifflin Co., 1937.

CHAPTER 6

GUIDED RESPONSE PROCEDURES¹

In the preceding chapter we discussed the use of free responses to test questions as means of measuring behavioral phenomena. We found that the scoring of such answers tended to be unreliable, and we presented various techniques for making them less so. Now, we shall turn to the body of procedures that have been developed in an effort to free educational measurement from the unreliability bug-a-boo, the true-false questions, multiple-choice items, matching questions, short answer, and fill-in items, and all the other guided response items that constitute objective tests.²

In our treatment, we first shall illustrate the many possible types of items and discuss briefly their basic attributes. Second, we shall examine the forms of measurement that we can hope to derive from guided response tests. After this we shall describe the construction of guided response instruments and their administration. Finally, the chapter will conclude with a short discussion of "standardized tests," this being the term ascribed to published guided response instruments for which norms are available.

¹ "Guided response procedures" is used as a categorical term for true-false, multiple-choice tests, and the like, in preference to the more customary "objective tests" for the following reasons. The term "objective tests" has no official or even semiofficial standing, as a technical phrase. It does not derive from any systematic classification of behavioral measurement procedures. The word "objective" can refer only to a single aspect of testing, that of scoring, while the construction and administration of such tests is as subjective a process as the construction and administration of any tests. Moreover, the phrase "objective tests" connotes (to the authors, at any rate) that the procedures are free from human error, which of course, they are not.

On the other hand, the term "guided response procedures" does derive from a systematic effort to classify the many activities and artifacts of behavioral measurement. It refers to what is thought to be the most salient characteristic of this approach to measurement, namely, restricted subject responses. And, finally, "guided response procedures" should not invoke any stereotyped connotation of infallibility.

² For a discussion of the history of such tests, see C. C. Ross and J. C. Stanley, *Measurement in Today's Schools*, New York: Prentice-Hall, Inc., 1954, chap. 2.

TYPES AND ATTRIBUTES OF ITEMS

The variety of test items used to elicit guided responses from pupils is very great.³ In this variety, though, it is possible to detect three basic categories of expected response. In many cases, pupils must select an answer from among several possible ones; in others, pupils must provide the answer themselves; and in still others, they are required to arrange words or objects into a proper array. We shall illustrate each of these three categories of items, and in so doing, describe the types of items most prevalent in guided response tests. It must be recognized that many others exist or can be devised and that combinations are possible and often desirable. Moreover, it should not be assumed that the precise form of any example is the only form for that type of item. Numerous variations are and can be practiced for each.

Selection of an Answer. Probably the greatest number of guided response tests employ questions for which pupils are expected to select the right answer from among those given. The appearance, the number of options, and the means of selection varies, but the principle of selection holds alike for true-false, multiple-choice, matching items, and all their variants. These are illustrated in Figure 9.

Example 2 in Figure 9 is similar to the other three in appearance, but it differs from them in its significance. The item is illustrative of the select-answer items used in personality and attitudinal measurement to gauge feelings.

Provision of an Answer. The items illustrated in Figure 10 require that a pupil provide his own answers, not just select one. Answers are limited to a word or phrase but, unlike the select-answer items, interpretation *does* play a small part in scoring. Notice Example 2 in Figure 10. In answer to the question, "What makes a cake rise?" pupils might respond with CO₂, carbon dioxide, gas, heat, baking powder, or soda. Are all these to be marked correct?

Because interpretation is involved (though far less than in free-response items) and because they may not be scored automatically, this type of item is not used in most standardized group tests. They are a mainstay of standardized individually administered tests, however, and they are possibly as widely used by teachers in their self-devised instruments as are the select-answer type.

Arrangement of Elements. The third basic category of guided response items illustrated in Figure 11 requires neither selection nor provision of an answer, but rather the arrangement of elements into a pattern. Of the three categories, these are the least used. Although scoring may be done with an exact key (no interpretation needed) the format of answers often precludes

³ A recent summary of research in achievement testing (9) states that there are no less than fifty different objective test techniques in common use.

1. *True-False*

Circle T or F.

T F In a city manager form of government, the mayor has little power.

2. *Affirm-Negate-Neutral*

Circle Yes, No, or U (Undecided).

Yes No U Minority groups should be more aggressive when faced with restrictive real estate covenants.

3. *Multiple-Choice*

Select the option that makes the statement correct.

- _____ Rainfall in Nevada is very light because
- a. Nevada has no large bodies of water.
 - b. The prevailing winds are northerly.
 - c. The Sierra Nevada Mountains screen the state from moisture-laden air.
 - d. Desert areas have a moisture evaporation rate too low for clouds to form.

4. *Matching*

Select the inventor from the right-hand column who goes with each thing in the left-hand column.

- | | |
|------------------------------|--------------|
| _____ 1. Cotton gin | a. Marconi |
| _____ 2. Electric starter | b. Colt |
| _____ 3. Steam engine | c. Howe |
| _____ 4. Wireless telegraphy | d. Edison |
| _____ 5. Atlantic cable | e. Kettering |
| _____ 6. Sewing machine | f. Field |
| | g. Franklin |
| | h. Whitney |
| | i. Watt |

Figure 9. Examples of guided response test items which require selection of an answer.

automatic scoring. Moreover, the items obviously are applicable only to subjects in which arrangement has particular significance.

Devices Employed by Guided Response Items

As you may surmise from the test items illustrated in Figures 9, 10, and 11, guided response procedures exploit all the basic devices by which measurement symbols may be assigned to behavioral phenomena. *Standard stimuli* are present in the questions themselves; it is assumed that all pupils perceive the same thing when they read the same item. In general, this assumption is warranted except for poor readers and foreign-speaking or culturally atypical

1. Fill-in or Completion

Write the correct word or number in each space.

The usual automobile wet cell battery contains _____ acid, and has a potential of _____ volts per cell.

2. Short Answer

Identify the following with a word or phrase.

The Duke and the Dauphin _____

Brom Bones _____

The Deacon's Masterpiece _____

Hiawatha _____

Answer the following with a number, word, or phrase.

What makes a cake rise? _____

What oven temperature is called *moderate*? _____

What quality do egg whites give to baked dishes? _____

3. Labeling

Label the parts of a living cell as shown in this diagram.

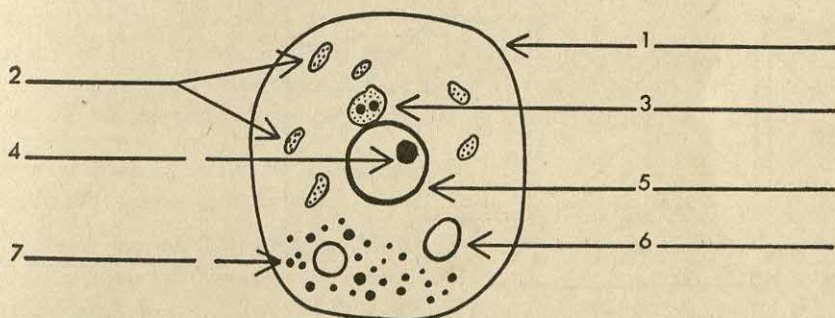


Figure 10. Examples of guided response items which require provision of an answer.

pupils. The limited number of possible answers, each with predetermined significance, makes for *standard responses* to the items. Guided response questions may be scored with a key, and thus they embody *standard analysis* in perhaps its purest form. Finally, as we will see later when we describe the construction of guided response tests, the test items *sample* that which they purport to measure. They do not, as a rule, survey the whole of it.

Advantages and Disadvantages of Guided Response Items

Guided response tests employ *standard stimuli* in common with free-response tests. They utilize *standard analysis* and must give attention to *sam-*

1. Ordering

Put these presidents in the proper chronological order.

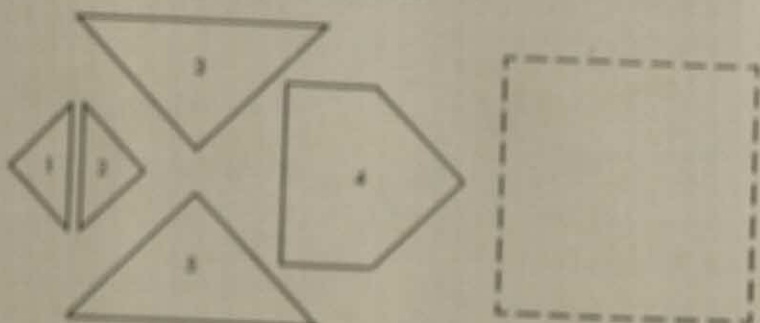
- a. Theodore Roosevelt
- b. Thomas Jefferson
- c. John Quincy Adams
- d. Benjamin Harrison
- e. William McKinley
- f. Abraham Lincoln
- g. Harry Truman
- h. Calvin Coolidge

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____

List these biological classifications in order from most general to most specific: Family, Genus, Variety, Phylum, Order, Species, Kingdom.

2. Assembly

Make a square out of these parts by sketching them in on the blank square and numbering them as they are numbered here.



Put these electrical symbols together so they could make an operating circuit. Draw in connecting wires as needed.

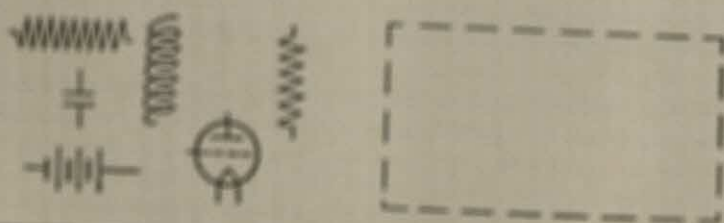


Figure 11. Examples of guided response time in which arrangement of elements is involved.

going along with all the basic procedures of behavioral measurement. Their unique attribute is that of *standard responses* and this gives the procedures their primary characteristic.

By having options of response and their meaning predetermined, it is possible to use a scoring system that involves little or no interpretation by the scorer and thus the subjective element in scoring can be eliminated or at least made negligible. It is possible, furthermore, to make comparisons among pupils on the basis of test scores since the test behavior of pupils is limited to a few options, each having the same meaning for all pupils. On the disadvantage side, limiting test behavior to a "yes" or "no," to a choice of a, b, c, or d, etc., automatically limits the pupil variation that the test can measure to the amount and kind of variation expressed in the options. Thus, where variation among pupils in a given subject is known to be greater than this, a guided response test may have limited validity. Moreover, *standardized responses* are best adapted to questions with exact and limited answers. In consequence, guided response tests can handle well the aspects of any subject that are exact as are the names and dates of history; but they are far less adequate for the indeterminate aspects, such as the causes, trends, and principles of history.

The items we have illustrated have been designed for written response. However, guided response procedures are as applicable to oral testing as they are to written testing. Many of the most valid intelligence and personality tests involve oral as well as written directions, and oral as well as written responses. In oral testing, the three types of items described are all to be found, and the four devices of standard stimulations, responses, analysis, and sampling are utilized just as they are in paper-and-pencil tests.

Forms of Measurement to Be Derived from Guided Response Test Scores

Raw Scores as Classification Numbers. The number of guided response items that a pupil answers correctly or in some given way is called a *raw score*.² This raw score, 10, 33, 72, 113, or anything else, is essentially a classification number with respect to the test. It says simply that on this test this pupil answered this number of items correctly. It is not a very precise classification number because two pupils might have the same raw score and yet have answered *different* questions correctly.

The fact that the raw scores on guided response tests are classification numbers is consistent with the function of the separate test items. Each item is designed to separate pupils into distinct categories. In tests of skill, knowledge, and intelligence the categories usually are two, those who answer the item correctly and those who cannot answer it correctly. In tests of personality and interest, items may yield three- or fourfold classifications but the function of the item still is only to classify. Any pupil's raw score then represents the number of times he is classified in a given way by the items of the test.

² In some cases it may be the number incorrect, but the significance is the same.

The Assumption of Equality Among Items. When a raw score is obtained in this manner by adding the correct or any other given category of item responses, the identity of the item responses is lost. Because of this, certain assumptions about the items must be made if the raw scores of different pupils are to be compared fairly. It must be assumed that all items have approximately the same difficulty or that their difficulty increases by even increments, that they have equivalent significance for the subject, and that all items have equal validity. To the degree that these assumptions are unwarranted, the same numerical scores may indicate different status and different scores might indicate equivalent status.

Conversion of Raw Scores into Rank and Scale Numbers. The classification numbers that are the raw scores may be converted into rank symbols, either plain rank or percentiles. If the test has had the validity, importance, and difficulty of its items established, the scores may under certain circumstances also be changed into scale numbers. The rank conversions may safely be made for raw scores obtained from the tests designed by teachers themselves, since rank differences imply no given differences in size or quality. Scale numbers, on the other hand, do imply such a condition and unless teacher-designed test items have been analyzed for difficulty, importance, and validity, raw scores may not be converted into scale scores. Statistical procedures for converting raw scores into rank and scale numbers are described in Chapter 7, pages 155-158.

THE CONSTRUCTION OF GUIDED RESPONSE TESTS

A guided response test is any collection of guided response items which, as a group, purports to measure something: knowledge of Spanish grammar, skill in copyreading, attitudes toward school, intelligence, personality structure, remembrance of the fifth chapter in Civics, anything. The test may be short or long, homogeneous as to type of item or heterogeneous, written or spoken, cover a week's study or a year's, and have a diagnostic or a survey purpose.

If the test is to be as valid, reliable, and efficient as possible in the light of what it measures and the purpose for measuring it, its construction is a complex and time-consuming task. Once a guided response test has been constructed, however, it may be used and reused with far greater economy of time than any other educational measuring procedure. Moreover, a great deal is known about guided response tests and one made in accord with this knowledge is likely to be an effective measuring instrument.

Some of this knowledge has come from systematic research, some from the reported experience of the users and designers of standardized tests, and still other from the practices of countless teachers.¹ The directions for the con-

¹ See Cook (3) for summaries of research on the construction of tests.

struction of guided response instruments, which follow, are derived from all these sources and as well from the experience and reflection of the writers.⁴

Definition of Phenomenon and Dimensions

Define exactly the phenomenon to be measured and its measurable dimensions in behavioral terms. The essential first step in preparing a guided response test, or a free response test for that matter, is to define exactly what you wish to measure and specify its dimensions in behavioral terms. This is in keeping with the conditions of measurability discussed in Chapter 2, which any phenomenon must approximate if it is to be measured.

To illustrate this first step in test construction, consider a teacher who wishes to appraise his pupils' knowledge of geography. He may begin by stating that "knowledge of geography" means "what pupils can remember, reason, and do that relates to the body of scientific knowledge about the earth, its climate, and the significance of each." Notice that the word "knowledge" now has been "spelled out" with words less abstract, that represent types of pupil behavior. Notice furthermore that "geography" has been reduced to elements more tangible and less general, and that the phrase, "body of scientific knowledge," defines the ultimate model of the content of any student's knowledge.

This matter of a "model" is particularly important for test construction. Obviously, if a pupil's understanding or knowledge of a subject is to be measured by a guided response instrument, there must be definite information about the possible content of the understanding or knowledge. Only this will enable the construction of test items that will properly sample the extent and depth of any pupil's knowledge. It is usual in testing for knowledge of subjects to assume that any pupil's knowledge approximates the body of organized knowledge that is the subject, however slight is the approximation and however erroneous it may be.

Another source of information about the content of pupils' understanding is the free writing and discussion of pupils. These may be collected and recorded over a period of time and from them may be outlined a "model" of knowledge or understanding. This procedure is applicable when the object of understanding of knowledge is not a defined and organized subject which is taught as such. Because, however, most school instruction relates to defined and organized subjects, this second method of determining the possible content of pupils' knowledge is little used by teachers.

The next step for the teacher is to have at hand or to obtain the facts and generalizations that constitute the body of scientific knowledge of the earth

⁴Detailed procedures for test validation and standardization are beyond the scope of this text. These are described in such books as *TREXER, Educational Measurement* (37), and *BEAN, Construction of Educational and Personnel Tests* (2). Techniques for finding an index of reliability and a standard error of score are presented in this text on pages 148, 181-186.

and climate which, by definition, must be the basis for the content of the test. Since the teacher has elected to measure "knowledge of geography" without any qualifications, the body of facts and generalizations on which he bases his test must be very comprehensive.

The ideal source of such facts and generalizations is all that has been published in scientific journals and books about earth and climate. The practical source for the teacher is a representative sample of these publications. This may include several geography textbooks and recent pertinent articles in geography journals. He may use a syllabus or course of study. It may simply be appropriate chapters in the textbook the pupils read, and/or his lecture-discussion notes for the course. As you may surmise, these sources are likely to provide progressively less adequate samples.

The use of a single pupil text or instructor's notes as the only source of content for a test of knowledge is somewhat justified if the test has only to do with understanding the text or the lectures and discussions. Examples of tests with such limited purposes are those of chapter or unit, or any designed to see what pupils have learned during a given instructional interval. However, even where the phenomenon to be measured is thus limited, there are serious deficiencies in the use of a single source. If the purpose is to see what pupils have learned, it is easily demonstrable that they always have learned more things, and things different from those the teacher and the text have presented. If the purpose is to see what has been gained from a chapter, restricting test questions to the content of the single chapter prevents measuring any applications pupils may have made of ideas in the chapter and any ideas having a dual source; for example, those that come in part from page 72 in the text and in part from an instructional film.

The problem of determining the factual content on which a test is to be based is peculiar to the measurement of subject achievement or aptitude, but it has a parallel in the measurement of attitudes, interests, and other attributes of personality. Just as it is necessary in measuring knowledge to have a definite idea of what the pupils may know, so in the measurement of personality attributes by guided response tests is it necessary to have beforehand some model of the probable feelings and behaviors of persons possessing varying degrees of the attribute in question. Often the principal sources of the model are the actions and feelings of those who:

1. Possess the attribute to an extreme degree and
2. Possess none or a minimum of it

In the design of vocational interest tests, the successful practitioners of given vocations often are the first group and those who have failed in them and those in antithetical callings are the second.

Now that the teacher has clearly defined knowledge of geography in behavioral terms and has specified its possible content, it remains for him to determine the dimensions he intends to measure. Some or most of the dimensions are suggested by the content itself. The facts, generalizations, ideas of

relationships, nomenclature, and the like that constitute the science of geography are the properties to some degree of any pupil's knowledge. Other significant dimensions are to be seen in the pupils' actions in relation to such content: the applications they make of it, the abstraction level(s) at which they operate, and the errors in their concepts.

Accordingly, the teacher in our illustration decides that he will measure each pupil's knowledge of geography with respect to

1. What and how many geographic terms he can identify (valley, cumulus, etc.) in each basic area of geography (sun-earth relations, maps, land features, winds, etc.).
2. What and how many basic facts he can state (maximum declination of the sun at the equator is $23\frac{1}{2}^{\circ}$, rocks in mountain streams are rounded, etc.) in each basic area.
3. What and how many generalizations he can make about causation, relationship, significance, etc., in each basic area (i.e., the energy of storms comes from the sun's heat).
4. What geographic skills he can perform (read and make maps, determine direction, etc.).
5. What applications he makes of his knowledge (i.e., says that Nevada probably never will have a very large population).
6. His errors in any of these above.

The construction of guided response instruments to measure achievement in other subjects or to measure personality, intelligence, etc., may require the designation of entirely different dimensions. The nature of these is discussed in Section II when the specific school applications of measurement and evaluation are discussed. However, it needs to be re-emphasized that one or more dimensions must be designated for the phenomenon before a guided response instrument may be constructed to measure it and that these dimensions must approximate certain conditions of measurability.⁷ As you may recall, the critical conditions for educational measurement are that the dimensions shall

⁷ A case for an exception to this rule might be made when so-called cut-and-try procedures of test construction are used. The gist of these procedures is to establish criterion groups and then to find test questions that discriminate between them most sharply without any particular concern over what dimensions the questions appraise. For example, if you wished to measure academic aptitude, you could select one hundred pupils whom teachers rated as excellent students and another one hundred whom teachers rated as very poor students, devise numerous test questions on any basis, and on any subject, administer them to the two groups, and select for your test of academic aptitude the questions which a high percentage of the excellent students and a low percentage of the poor students answered correctly. In this procedure, however, the originally devised questions necessarily reflect some tacit, and perhaps unconscious, hypotheses about the dimensions of academic aptitude. They certainly were not drawn at random from a barrel labeled "Test Questions, Miscellaneous, Unsorted." Consequently, we propose that dimensions be explicitly designated prior to test construction, even if empirical findings on item discrimination are to be the basis for test revision. It is thought that the great majority of test theorists agree with this position. See, in particular, Flanagan (14), and Travers (37).

1. Provide sensory data
2. Be clearly defined
3. Produce consensus among impartial and unrelated observers

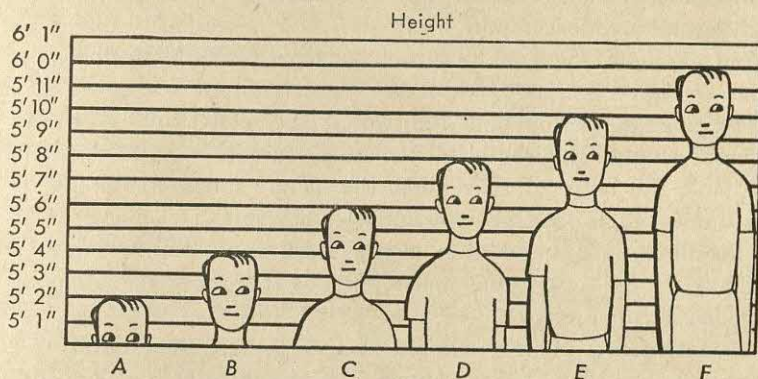
In addition, if pupil achievement is to be evaluated as a result of a guided response test, dimensions should be selected in view of the standard on which the evaluation is to be based. More often than not, a guided response instrument is used not merely to measure a pupil's status but also to evaluate it. Should a performance scale be the evaluative standard, it is essential that test items be keyed to each of the several levels of the standard. For example, if the standard contains a level of evaluation and interpretation of facts as distinct from simple knowledge of them, the dimension evaluation and interpretation must be measured by the test as well as the several dimensions of factual knowledge. (See Chapter 9 for a full discussion of evaluative standards and their implications for measurement.)

Preparation of Items

Devise verbal and/or graphic items, the responses to which automatically will classify pupils with respect to given dimensions. As we have seen, a single guided response item is capable only of classification and, in most instances, twofold classification. Pupils usually will show many degrees of difference for any measurable dimension and total test scores may be considered to reflect this "continuous" variation. However, so far as item-by-item responses are concerned, the continuous variation has been broken down into a series of all-or-none variations.⁸ For example, take the first dimension that the geography teacher in our illustration wishes to measure, "What and how many geographic terms a pupil can identify." Pupils may be expected to vary from those who know practically none of the terms through those who know intermediate amounts to the few who know a great number. No two pupils, in actuality, are apt to know the same ones or the same amount. But when the pupils respond to a question keyed to this dimension (i.e., cumulus means a *glacier*, a *fossil*, a *cloud*, or a *rock formation*), they will be separated into two groups, those who answer "cloud" and those who give any other answer, "fossil," "glacier," "rock formation," or no answer at all.

The rationale that permits us to obtain a measurement of a many-degreed or "continuous" dimension through a series of twofold classifications is made graphic in Figure 12. In these illustrations, notice that many degrees of variation are possible relative to the lines on the wall (representing height) or to the grid (representing knowledge of geographic names) but that only two degrees of variation are possible for each line or each section of the grid. A number that represents the height of each of the six boys is obtained by count-

⁸ In the measurement of personality traits and structure and, in some cases, achievement and intelligence, item responses may be used for tripart or even four-part classification.



Boy *A* is 5' 2", *B*—5' 4", *C*—5' 6", etc. Their height can be computed by making a yes or no classification for each with respect to each line on the wall, thus: *A* is yes for 5' 1" and 5' 2", but no for the others; *B* is yes for 5' 1", 5' 2", 5' 3", and 5' 4", but no for the others; *C* is yes for 5' 1" through 5' 6", but no for the balance; etc., until *F* is yes for all but 6' 1".

Knowledge of Geographic Names

1	7	13	19	25B	31
A	8	14	20	26	32
2	C	15	21	27	33
3	9	16	22	28	34
4	10	17	23	29	35
5	11	18	24	30	36
6	12	19	25	31	

Let each of the 36 sections of this grid represent knowledge of one geographical name, and the area covered by *A*, *B*, and *C* respectively indicate the extent to which three pupils know these names. The area and shape of *A*, *B*, and *C* could have been determined by making a yes or no classification for each with respect to each of the 36 sections. *A* is yes to 8–29 or 16 sections, but no to 1–7 and 30–36. *B* is yes to 21, 22, 26, 27, 28, 33, 34, and 35 or 8 sections, and no to 1–20, 23, 24, 25, 29, 30, 31, 32, and 36. *C* is yes only to 15, 16, 21, and 22, or 4 sections, and no to all the rest.

So 16 represents *A*'s knowledge of names, 8 represents *B*'s, and 4 represents *C*'s.

Figure 12. Measurement of many-degreed dimensions by a series of dichotomous classifications.

ing all the inch lines touched by each boy. Similarly, a number representing the geographic knowledge of each pupil may be found by counting all the name squares that each pupil covers.⁹ For this process to be valid, it is necessary to

⁹ The rationale underlying tests employing items that make more than twofold classifications is simply an extension of the rationale for twofold items.

assume that height and knowledge of geographic names may be defined as a group of elements each of which may be possessed or not possessed and all of which have equivalent importance.¹⁰

In our example, then, the geography teacher's task is to make up items which, if answered correctly, will classify the pupil as possessing given elements of a given dimension, and, if answered incorrectly, will classify the pupil as not possessing those given elements. As a source of the items, the teacher has available the books, the syllabus, the lecture notes containing the geographic names, facts, and generalizations that are the elements of the first four of his dimensions. In addition, he may need to use pupil products and his recollection of pupil discussions and reports as sources of items having to do with applications and errors.

The construction of items should be governed by the following general considerations and rules. The advantages and disadvantages of items of different types will be discussed later and specifications given for each type.

FORM

As a rule, guided response items should be short, affirmative, simple and, of course, unequivocal. They need to be short so that the significance of items may be grasped by slow thinkers and poor readers as well as by fast thinkers and good readers. If the idea to which the item refers is complex or if intelligence or reading ability is being measured, this rule may need to be excepted. Negative statements usually hinge upon a single word, *not, no, can't, didn't*, etc. If the reader fails to notice the word, he will fail to read the item properly. In an affirmative construction, on the other hand, meaning is not so dependent on a single word. If a negative construction has to be used, the negative word should be underlined.

Items whose syntax is simple are preferable to those whose syntax is complicated. Unless facility in reading complex sentences is being tested, the correctness of a pupil's response should not depend on his ability to solve a grammatical puzzle. Finally, it is a truism in guided response testing that each item and each response option must have *just one* meaning. Ambiguity of items is a principal cause of unreliability in tests.

LANGUAGE

Items should be couched in pupil language. A pupil's knowledge consists of what he thinks and says and writes. His attitudes consist of feelings toward things and ideas as he conceives of these things and ideas. When we measure these by a guided response technique, we assume that a given response to an

¹⁰ The term "elements" is meaningful for the guided response measurement of most dimensions of educational significance, but is a misnomer in some cases, e.g., *reading comprehension, intensity*, etc. Here, a more appropriate phrase might be "a function of many single responses that occur or do not occur."

item means that the pupil thinks or feels in that given way. Consequently, it is essential that both items and answer options be phrased in language like the language pupils use in thinking and talking. Only if we do this can we be sure that the pupil's response is on the basis of what he knows or feels. If the language of the item is too strange, too bookish, or simply too difficult, we do not know the basis on which any pupil responds; it may be guess, intelligence, misinterpretation, etc., but we cannot assume that it is on the basis of his knowledge or attitudes.

This does not mean that guided response questions should avoid the use of technical terms, "big words," and precise grammar. They should use technical or academic language whenever this is the way pupils must write and talk if they know the subject. It is the unnecessary use of these that is to be avoided. In the following example, an item intended for an eighth-grade test but written sentimentiously is translated into more appropriate language.

- T F The Senate's failure to approve the League of Nations constituted a censure of Wilson's relations with Congressional leaders.
- T F The Senate failed to approve the League of Nations because Wilson had slighted some of its leaders.

LIFTED ITEMS

Do not "lift" items from text or reference books unless the object of the item is to measure rote memory of what has been read. Except for the axioms and formulas of mathematics and certain "laws" or generalizations in the sciences, the language of pupils' knowledge is not likely to be just like the language of textbooks. So if pupil language is to prevail, the use of the exact wording of sentences in textbooks is generally to be avoided. Moreover, even if the language of the book is simple, "lifted" questions tend to measure rote memory of book statements. In particular is this true in completion or fill-in items in which the context of a statement determines its key words as much as the statement itself. To illustrate this point, consider the following fill-in item, which consists of a statement taken verbatim from a World History text with three words omitted.

None of the three _____, _____, or _____ especially encouraged the building of civilizations.

According to the context (religion, philosophy, or geography) a correct response might be Buddhism, Christianity, Taoism. It might be Thoreau, Adams, Gandhi. Or it might be (what it is intended to be) highlands, forests, grasslands.¹¹ To be able to answer the question correctly, a pupil would have to remember that this exact statement occurred in the book.

¹¹ F. C. Lane *et al.*, *The World's History*, New York, Harcourt, Brace, and Co., 1954, p. 23.

SCORING

Prepare for responses to items to be given in the simplest way possible and, if feasible, so as to permit machine or stencil scoring. In general, for select-answer items it is considered better to have options checked or circled than it is to have them written in. Time is saved and there is less likelihood of indeterminate responses. Scoring is facilitated by having all responses occur in a straight column at the left or right margin. If an electrical scoring machine is available, select-an-answer items may be answered on separate machine-scoring sheets with resulting economy of time and increased accuracy in scoring. Standard score sheets may be obtained from the manufacturer of the machine or from test publishers. Stencil scoring is a mechanical equivalent of electrical scoring and may be used with a separate score sheet or with the test itself as long as options are checked, crossed out, or circled. Stencil scoring means that a cardboard or piece of heavy paper has had a hole punched in the position of each correct response. The stencil then is laid over a student's test paper and those marked are counted.

Arrangement of elements and provide-an-answer items do not as readily lend themselves to a uniform response format. However, an effort should be made to have responses to such items occur toward either margin and in approximately the same place for successive items.

DISCRIMINATION

Devise items that will discriminate among all types of pupils to be tested and include only discriminating items. The function of a guided response item is to classify pupils relative to some dimension. So it follows that the item must discriminate among the pupils. If all pupils answer it in the same way, it does not classify them and the item has measured nothing. Therefore, no items should be devised that are so easy that all pupils will answer them correctly, or so difficult that all will answer them incorrectly, *when the purpose of the item is to measure*. If an item is meant to give confidence to certain pupils or to deflate the egos of others, it may be used on motivational grounds but not on the grounds of measurement. Responses to items of this sort should be excluded from total scores.

An apparent exception may be made to this rule when there are specific instructional objectives that all pupils must achieve and to the same degree. For example, in beginning algebra all pupils must learn to transpose terms ($x = y$ may be stated $x - y = 0$). A test in algebra might have an item keyed to such an element in full expectation that all pupils in a given class would answer it correctly. In a larger population of pupils, however, there would be some (those who had not studied algebra) who would answer incorrectly. Hence, the item would discriminate with reference to this larger population.

Not only must test items discriminate, they must *discriminate positively*. For achievement and intelligence tests positive discrimination means that a

greater proportion of pupils who get high scores on the test answer the item correctly than pupils who receive low scores on the test. A total score for a test must bear a direct and not an inverse relationship to any item score. Obviously, this is essential if the size of a total score is to be directly related to the degree to which a pupil knows or feels a given thing. The degree of positive discrimination for items often is used as a statistical index of their validity.

A simple way to check items both for general discrimination and positive discrimination is to pick papers whose total scores represent the top 25 per cent and those whose scores constitute the bottom 25 per cent.¹² Then for each item tabulate the percentage of each of these groups who answered the item correctly. Figure 13 shows a section of a table resulting from this item analysis process.

<i>Item</i>	<i>Per cent of upper quarter who answered correctly</i>	<i>Per cent of lower quarter who answered correctly</i>
23	75	35
24	50	45
25	100	100
26	25	40
27	40	10

Figure 13. Section of an item analysis table.

With reference to Figure 13, all items discriminate except Number 25. It would be necessary to check all the papers to determine whether 25 has any discriminating power at all. Items 23 and 27 show marked positive discrimination. Item 24 shows a slight degree of positive discrimination but Item 26 shows negative discrimination. Consequently, items 25 and 26 are suspect until proved valid or are revised; items 23 and 27 are clearly good items; while Item 24 should be inspected for flaws but may be valid.

Item analysis is an empirical process requiring at least one administration of the test items. It is useful generally for tests or test items that are to be reused. The analysis gains significance if several administrations of the items contribute to the figures. One convenient mechanical arrangement for item analysis is to write each item on a separate card and to enter there a tally of pupil responses for each administration of the item (see Figure 14). Then, in immediate view of the results over a period of time, the item may be inspected for discrimination and difficulty.

While item analysis is time-consuming, it is the only way known to *verify* the validity of items. Its continuous use with a large inventory of items placed

¹² Additional statistical devices for item validation are described in such texts as Bean (2) and Tate (34).

<i>The largest mountain range in the United States is (a) Sierra Nevadas, (b) Rocky Mountains, (c) Appalachian Mountains, (d) Cascade Mountains.</i>				
<i>Date</i>	<i>Number of pupils</i>	<i>Number answering correctly</i>	<i>Per cent correct in 4th quarter</i>	<i>Per cent correct in 1st quarter</i>
12-56	42	30	74	58
12-55	45	35	80	63

Figure 14. Illustration of a test item entered on a card together with an analysis of responses to the item.

on separate cards facilitates the construction of any particular test on the same subject. The necessary type and number of items may simply be selected from the card file, arranged in proper sequence, and typed for duplication. New items must, of course, be designed as those that prove to be invalid are discarded and as new dimensions or subject elements are to be tested.

As a rule, it may be assumed that items you think will discriminate positively actually will do so. Technical causes of no discrimination or negative discrimination in items have been summarized by Lindquist (22) much as follows: (a) weaknesses in the item: ambiguity, clues, misleading elements, etc.; (b) insufficient learning of the conception to which the item is keyed, so that a plausible false option appeals to superior pupils; (c) a decoy response that appeals to conflicting learning from other subjects, to conflicting prejudice, or to peer opinion, with the right response appealing only to experts.

DIFFICULTY

If technical flaws in items or improper test conditions do not produce the varying degrees of discrimination observed for achievement and intelligence test items, it is assumed that the relative difficulty of the items produces the effect. Thus, degree of discrimination may be taken as an index of item difficulty. An item which 90 per cent of a group answered correctly would be considered an easy item. One which only 10 per cent answered would be termed very difficult. An item that half of a class answered correctly and half answered incorrectly is said to have 50 per cent difficulty.

The question of just *how* difficult should be the items in a test has been studied for many years but it remains a moot point (9). Perhaps, what is an optimum test may not be stated without reference to a particular measurement purpose, and, therefore, neither may optimum item difficulty be discussed in the abstract. Obviously, tests used to select a small number from among a great number (a foreign service examination, for example) may include items of greater mean difficulty than tests designed simply to measure variation among an unselected population (for instance, an adult group intelligence test).

A test for speed of performance should, of course, have items of the same difficulty, whether this is high or low. In power tests, those designed to measure the level of a pupil's ability or knowledge, it is usual to have items that increase in difficulty as the test goes on. The last question a pupil can answer correctly is taken to be indicative of his level of achievement. Hence, items should be carefully graded as to difficulty and placed in ascending order of difficulty. Power tests are used widely in the measurement of intelligence, mathematics, vocabulary, and reading.

In tests directed at achievement in general, particularly toward what a pupil knows, how much he knows, and how accurate is his knowledge, we advise that items be constructed without special regard to their difficulty. Careful attention should be given to adequate sampling of the dimensions of the subject. If the sampling is adequate, the items will vary in difficulty according to the gradations of difficulty inherent in aspects of the subject. If there are no gradations of difficulty among elements of the subject, then item difficulty is not a relevant consideration.

Item difficulty not only is a function of the element to which it is keyed but also of the type of item. This type difference in difficulty is not usually included as a factor in analyses of item difficulty based on their discriminating power, and only items of the same type should be so compared. The inherent difficulty of a given type of item may be a factor in selecting that type of item for use; it may affect the weighting of items; but it should not be confused with the difficulty of the knowledge element to which the item is keyed.

The foregoing discussion of item discrimination and difficulty has related to tests of ability or subject achievement. The basis for discrimination in tests of personality, attitudes, and interests will differ, but the need for positive discrimination in test items is as applicable to the measurement of personality variables as it is to achievement. Item difficulty has, of course, no relevance for testing in these areas.

INDEPENDENCE

Items should be mutually independent. The usual interpretation of scores on guided response tests assumes that a response to one item is not affected by a response to any other item. Without independence of items, the probability factor in guided response items becomes complex and tends to distort the relationship between scores and achievement. A pupil who knows the item that is a clue to another has an advantage over the pupil who does not.

For example, suppose that in a six-item test the probability of a correct response to items 2 and 3 is doubled if there has been a correct response to Item 1 and it is reduced to zero if there has been an incorrect response to Item 1. On the other hand, the probability of a correct response to items 4, 5, and 6 is not affected by a correct response to Item 1. Two pupils, each with an ability to get 50 per cent of the items correct, might make different scores just because of this condition, as follows:

FUNDAMENTAL CONCEPTIONS AND PROCEDURES

Item	Independent probability of a correct answer to any item for either pupil	Probability of Pupil A having correct answers if he answers 1 incorrectly.	Probability of Pupil B having correct answers if he answers 1 correctly.
1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
2	$\frac{1}{2}$	0	1
3	$\frac{1}{2}$	0	1
4	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
5	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
6	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
Most probable total score	3	2	4

Mutual independence does not mean that several items may not be keyed to the same element. It means simply that the answer to one item should not be stated or suggested in the body of another or that the significance of one item should not be contingent upon a proper answer to another. Items 17 and 35-39 below illustrate the "clue type" of interdependence, and items 43 and 44 the "contingency type."

NOTE. These items are bad examples.

- T F 17. The Columbia River empties into the ocean in the state of Washington.

Name five navigable United States rivers that have industrial and hydroelectric importance.

35. _____
 36. _____
 37. _____
 38. _____
 39. _____

43. What kind of clouds cause thunderstorms? _____
 44. The usual mean depth of these clouds is
 a. 3,000 feet
 b. 6,000 feet
 c. 500 feet
 d. 1,500 feet
 e. None of these

SAMPLING

As a group, items should sample adequately all dimensions subject to sampling of the phenomenon being measured. Unfortunately, there are no

precise directions that may be set forth to insure proper sampling by items. We have discussed earlier the general nature of sampling and some principles that pertain to it (pages 38–40). These principles are as applicable to this sampling task as to any other.

The number of parts and/or subdivisions involved in a dimension somewhat govern the number of items that should be keyed to it. In illustration of this, consider the first dimension the teacher in our example designated for knowledge of geography, "What and how many geographic terms the pupil can identify in each basic area of geography." If there were three basic areas, earth, sea, and climate, and each involved 100 terms, 25 questions might be keyed to each area. On the other hand, if there were 1,000 terms for each of the areas, sampling would seem to demand more items for an equally good sample. If the areas were unequal as to the number of terms they subsumed, the number of items keyed to each area should be somewhat proportional to the different number of terms in the areas. However, there is no exact relationship between the number of elements in a dimension and the adequacy of given size samples. Just because 1,000 geographic terms might relate to the earth and 500 to the climate, a sample of 50 earth-keyed items would not necessarily be equivalent in adequacy to 25 climate-keyed ones.

Aside from the "how many items" aspect of sampling, it is necessary to consider what items are needed to give an adequate sample. The rule, of course, is that the sample exactly represents the whole, just as a reduced photograph exactly represents a larger one. This condition may be approached, first, by including items for each subdivision or aspect of a dimension. If the number of components of the subdivisions is known, the number of items for each should, as we have stated, be proportional.

Second, it is necessary to select the actual elements to which items are to be keyed so that each element in the subdivision has equal probability of being in the sample. For example, if recognition of names of navigable rivers is to be measured, each navigable river should have as good a chance as any other navigable river of being in the test. To accomplish this, it is necessary to use some system of random selection. Pages may be chosen in an automatic fashion, every fifth, tenth, etc., from text and reference books, or facts and ideas may be written on slips and a sample literally drawn from a hat.

An objection frequently raised to a random selection of elements is that some elements are so important that they must be included. If this is the case, then these very important items should constitute a separate subdivision to be included *in toto* or to be sampled separately. The device of random selection is intended only for and is only valid for groups of homogeneous elements of equivalent importance.

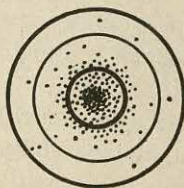
An empirical procedure is available to determine when items are sufficient in number and nature to sample a dimension properly. Use a given number of items keyed to a given dimension in a test, administer the test, and note the results. Add items keyed to the same dimension to the test, readminister it to the same group, and compare the results with the first testing. Continue this

add-items-and-retest process until the continued addition of items produces no change in the relative standing of pupils on the test. The number of items you have at this time is all you need. The procedure, regrettably, is time-consuming and difficult and usually is employed only in the construction of standardized tests.

Types of Items

Select types of items for use according to their special properties and construct them according to their type specifications. The characteristics of the three basic categories of guided response items (select-an-answer, provide-an-answer, arrange elements) as to their chance factor, administration time, and difficulty are summarized for comparative purposes in Table 2.

Chance Factor and Administration Time. From this overview it is apparent that select-an-answer items are the most efficient for time of administration, but are likely to be the least reliable with their high chance factor. The chance factor means simply that by guessing alone a pupil would be expected to make an indicated percentage of correct responses. With chance a factor in determining pupils' scores, the reliability of those scores is reduced since a proportion of any score must fluctuate in a random fashion. This situation is analogous to the "scatter" found for a group of shots from a rifle. Even though the rifle is clamped in a vise and has been "zeroed in" on a target, not all the shots hit the same spot, but will describe a cluster thus:



with the position of any given shot partially determined by *chance variations* in ammunition, wind, barrel whip, etc. The larger these variations the larger the cluster or the greater the unreliability of the shooting. Just so, larger chance factors make for greater inherent unreliability in test items.

Provide-an-answer items, on the other hand, have an advantage so far as inherent reliability is concerned. Chance may be considered a negligible factor. But they are at a disadvantage relative to time of administration. No data are available as to the administration time of arrange-elements items, but their inherent reliability in most cases should be intermediate between that of provide-an-answer and select-an-answer types.

Difficulty of the Three Types of Items. The relative difficulty of the three types of items, if subject difficulty is the same, is a function of the psychological activity involved in responding to them. To answer true-false, multiple-choice, and matching questions correctly requires merely that a right response

TABLE 2

Types of Guided Response Items Compared as to Chance Factor, Time for Administration, and Difficulty

Type of item *	Chance factor ‡	Items administrable in ten minutes †				Relative difficulty	
<i>Select-an-answer</i>							
True-False	50%	Elementary 42	Low 10 per cent 36	Secondary Average 54	Upper 10 per cent 72	College 71	Least difficult because of appreciable chance factor and the fact that only recognition of a correct response is required.
Multiple-Choice	33%		25	37	50		
3 option	25%		(Average for 3-6 options)				
4 option	20%						
5 option							
Matching	1st item = $\frac{1}{\text{No. in selection column}}$ 2nd item = $\frac{1}{N-1}$ 3rd item = $\frac{1}{N-2}$ etc.	No data, but probably approximates multiple-choice.					
<i>Provide-an-answer</i>							
All types	Indeterminate but may be considered as negligible	20	20	30	40	33	Most difficult because no appreciable chance factor, and recall of a correct response is required.
<i>Arrange elements</i>							
All types	A function of the number of elements, and the type of arrangement required.	No data available					Difficulty is a function of the individual item, can be little or great.

* For examples, see figures 9, 10 and 11.

† Adapted from Tieg, (36, pp. 93-95). The figures for secondary pupils are the results of a specific piece of research while the other figures are composites.

‡ The method of computing the chance factor for matching items is only approximate since it assumes that the correct response for any given item has not been chosen incorrectly for a previous item.

be *recognized*. This is tantamount to recalling or "sensing" that a present perception has been perceived previously. As we know, this represents a simple and easily attained learning of which even lower animals are capable.

Provide-an-answer items are likely to be more difficult since verbal recall is a more exacting psychological activity than recognition. In general terms, "recall" means that part or all of a previous experience may be re-enacted in response to some fraction of the stimulatory field that produced the original action. Higher animals are distinguished by their facility for such re-enactments and it appears later in the mental development of the human infant than does facility for recognition.¹³

It is likely that arrange-elements items have about as much inherent difficulty as provide-an-answer ones.¹⁴ Although recognition of elements is involved rather than recall, recall of a pattern or reasoning out of a pattern is necessary. Moreover, since arrangement items can be keyed to very complex patterns of relationship, they may have a greater difficulty potentiality than any of the other types of items.

In addition to the relative difficulty, reliability, and administration time of types of items, it is necessary to consider some of their unique characteristics, their usual applications, and specifications for constructing them.

SELECT-AN-ANSWER ITEMS AS A GROUP

True-false, multiple-choice, and matching items are familiar to pupils, their chance factor is exact and they are the easiest to score of any items. Only select-an-answer items are currently amenable to machine scoring. They are best adapted to measuring areas of knowledge or ability that involve many specific items of exact information or performance. Agree-disagree or yes-no-maybe variants of true-false items and the others are, in addition, useful for gross screening in the measurement of personality attributes. Select-an-answer items have been found of less use in such subjects as art, music, physical education, shop, etc., and in appraising those phenomena we call skills (writing, drawing, studying, etc.). In subjects in which relationships and patterns are significant, and facts per se are less so, select-an-answer items have limited usefulness, philosophy, sociology, and psychology, for example.

A frequently voiced criticism of select-an-answer items and one with which the authors must agree is that they are artificial in character. It is axiomatic in measuring behavioral phenomena that the test situation is best which approximates most closely the actual situation in which the phenomenon being measured usually occurs. Now, thinking true or false to specific questions, mentally reviewing optional answers to a specific question and then selecting one, or matching up columns of associated ideas does not describe the

¹³ The actual distinction between recognition and recall is considered to be one of degree rather than kind by most psychologists.

¹⁴ The writers are aware of no research bearing on their difficulty and the psychological activity in responding to them is not as clear-cut as it is for the other types.

anatomy of knowledge, personality, intelligence, or anything but behavior during a test.

TRUE-FALSE ITEMS

While they are the most economical of time and space, true-false items are particularly susceptible to stereotyped construction and, thus, to the operation of syntactical clues. One early study of the syntax of true-false items (5) found that longer items were more likely to be true; four of five statements containing *all* were false; three of four containing *always* and *never* were false; four of five containing *no*, *none*, or *nothing* were false; and nine of ten using the word *only* or *alone* were false. For some reason, stereotyped expressions tend to accompany false items and other stereotyped expressions tend to accompany true items. Pupils become habituated to these stereotypes and try to answer many questions on the basis of their form. A second general liability of the type is that true statements must be totally true while false ones may be only partially false. For this reason, true statements not at the same time give-aways are sometimes very difficult to construct.

Whatever their liabilities though, well-designed true-false items have great usefulness in educational measurement. Among the specifications for their construction are the following.

1. Avoid the expressions that tend to go with false items: all, always, never, no, none, nothing, only, and alone.
2. Make true and false items of about the same mean length.
3. Avoid statements so abstract or so general as to be unanswerable by a true or false, or yes or no.
4. Have truth or falseness be a function of the basic predication of a statement, not of some minor element, clause, or phrase. For example.

(Right)

T F Calvin Coolidge was famous for his reluctance to speak at length.

(Wrong)

T F Calvin Coolidge, elected president in 1922, was famous for his reluctance to speak at length.

In the first item, truth or falseness properly hinges on Coolidge's taciturnity and is true. In the second, while the basic predication of the statement is still true, the phrase "elected president in 1922," is false since there was no presidential election in 1922. Hence, the statement as a whole is false.

5. Avoid such vague qualifiers as usually, seldom, much, little, many, few, large, small, etc.

6. Make approximately half the items true and half false and determine the order of true, false by chance.

7. Have *enough* true-false items so that the chance factor (50 per cent) does not too seriously restrict the range of the test or the reliability of a score.

Ten items are certainly too few¹⁵ for five could be answered correctly by guess alone and, thus, the effectual range of scores is only five. In such a restricted range chance would be a disproportionately large factor in any pupil score. We suggest that fifty true-false items be considered a rule-of-thumb minimum.¹⁶

MULTIPLE-CHOICE ITEMS

Those items offering 3, 4, 5, or even 6 possible answers from which the testee is asked to choose (see Figure 9) are more difficult to construct and less economical of space than true-false items. The difficulty in their construction lies chiefly in designing plausible decoys or false options. On the advantage side, their chance factor is less than that of true-false items, and they can be keyed to more complex elements of knowledge or feeling than can true-false items. It sometimes is advantageous to combine multiple-choice and true-false types. For example,

(Mark each option true or false)

The Mississippi River affects the economy of the United States as follows:

- | | | |
|----------|----------|--|
| T | F | a. It is used for passenger traffic. |
| T | F | b. Coal barges sail on it. |
| T | F | c. It serves to irrigate Iowa and Illinois. |
| T | F | d. It produces more hydroelectric power than any other American river. |
| T | F | e. Its immediate valley is rich farming country. |
| T | F | f. It is navigable as far north as Rock Island. |

The following specifications are offered for the design of multiple-choice items.

1. Four or five options are standard practice and the use of at least four is advised. With three options, the chance factor is nearly as great as in true-false items.

2. The number of multiple-choice items needed for reasonable reliability and range of scores is, of course, less than for true-false items. If we use an effectual score range of 25 as a rule-of-thumb minimum, we should have at least 33 four-option items, 31 five-option, or 30 six-option. Chance would, on the average, account for about 8 points, 6 points, and 5 points respectively of any pupil's total score.¹⁷

3. No options may be absurd or obviously true. Absurd options reduce

¹⁵ If a quiz (as on a reading passage) is used for motivational purposes only or just to see who have read the passage and who have not, number of items is not such an important consideration.

¹⁶ Assuming that pupils will guess at some answers. See pages 105, 118.

¹⁷ This assumes that pupils will guess at some answers. See pages 105, 118.

the number of options that test anything by the number that are absurd. Obviously true ones permit all pupils to select the same one and thus the item will not discriminate.

4. If a negative statement is used, underline the negative word so that it will not be overlooked.

5. It is advisable to have options be the responses to a question, the solutions to a problem, or the predicate of a sentence whose subject is the basis of the item. Such simple constructions are likely to offer fewer contextual clues to the correct answer and the effect of reading skill and intelligence hence are minimized.

6. All options should be grammatically consistent so that the syntax of the item will neither help nor hinder a correct response.

7. Obviously, the more nearly true or plausible decoy options are, the more difficult it is for the pupil to detect the correct response. Hence, the near-correctness or plausibility of options may be increased to measure higher levels of knowledge or intelligence.

8. Correct options must not always be in the same place, nor should they fall into a pattern. Chance selection of the correct position will insure both conditions. A table for use in randomizing multiple-choice options has been prepared by Anderson (1).

MATCHING ITEMS

Matching items are the easiest of all to construct and are economical of space. However, their use largely is limited to associative pairs: presidents and dates, wars and dates, authors and books, stories and characters, inventors and inventions, etc. Adherence to the following rules will insure maximum reliability in their construction. (For example see Figure 9).

1. The best number of items in the stimulus column (things with which others are to be matched) has not been precisely determined by research, but probably is between 5 and 12.

2. There should be more items in the response column than in the stimulus column (say three more) so that the last of a series may not be matched just by elimination.

3. Stimulus and response columns should be on the same page, side by side, with the stimulus column to the left.

4. Directions should clearly state which column is to be matched with which, and what is to be written in, letters, numbers, or words.

5. A single matching series should involve a single subject (wars-dates, books-authors, etc.) and not several subjects. If heterogeneous subjects are included, correct answers may be gained by reasoning as well as by knowledge.

6. Items in one of the two columns may be listed in some logical order, but items in the other must have a random sequence so that item position can be no clue to that which it matches. Random sequence (so far as the significance of items is concerned) frequently may be obtained just by putting the items in alphabetical order.

PROVIDE-AN-ANSWER ITEMS AS A GROUP

Completion, short-answer, and labeling items are illustrated in Figure 10. As a group, they are a more natural type of item than select-an-answer. Their form more nearly approximates the way subjects are presented, and responses to them are more like the actions and thoughts that we say constitute subject achievement. Like the other category, though, they usually are applicable only to subjects in which items of exact information and/or precise statements are important. Labeling items are particularly useful in scientific and mechanical courses.

The construction of provide-an-answer items is relatively easy, but scoring them is more tedious than select-an-answer. They may not be machine scored. Frequently, more than one exact answer may be correct and some interpretation on the part of the scorer is necessary.¹⁸

COMPLETION OR FILL-IN ITEMS

Completion items are easy to prepare and can be used to measure retention of a composite idea. For example, pupils' knowledge of an entire electrical concept may be tested by omitting the underlined words in the following sentence.

In electricity, the resistance of a wire determines the amount of current that will flow in response to a given amount of voltage.

In constructing completion items, the following procedures are advisable.

1. Omit *important* words or short phrases only.
2. Ascertain that only one word or short phrase is the correct response. If this is impossible insure that only a limited number of insertions will be correct and that you know each of them.
3. Do not leave so many blanks in a given statement that the statement loses its meaning.

¹⁸ An absolute distinction is impossible between provide-an-answer items, which we classify as guided response, and those which we classify as free response. As short-answer responses become longer, and as free responses become shorter, the two must converge, of course. As a basis for distinguishing between the two types, we offer the following:

- a. If necessary, select-an-answer items could be substituted for provide-an-answer ones. They could not be substituted for free-response items.
- b. The answers to a provide-an-answer question will be either right or wrong, will get full value or no value. The answers to a free-response question will, on the other hand, range from those entirely wrong through those of little, of medium, and of great value to those entirely correct.

Distinguishing between the two types is not entirely an academic point. Free-response items require a different sort of scoring key from guided response items, weighting has a different significance for them, and they can easily appraise aspects of knowledge that guided response items can tap only with great difficulty.

4. Make all spaces in the same statement the same length, so that the size of space can be no clue to the length of the right word or phrase. Any space should be large enough for the longest word or phrase.

5. Within a single statement, omitted words or phrases should be of parallel grammatical significance to prevent the operation of syntactical clues.

SHORT ANSWER ITEMS

Since responses may be arranged to occur in a vertical column, short-answer items may be easier to score than the completion or fill-in type. On the other hand, they do not lend themselves so well to testing complex ideas since only one response usually is made to each question or problem. Other than these differences, they are similar to completion items and their construction should be guided by two of the same maxims: only one word or phrase is the correct answer; and space for responding should be no clue to the length of response. In addition:

1. Items should be direct questions or commands, as,

What is the formula for sulphuric acid? _____
or

Give the formula for sulphuric acid. _____

2. Questions or commands should be directed to the important aspect(s) of a fact or idea rather than to superficial or trivial ones, as,

Clements' pseudonym, "Mark Twain," derived from what Mississippi steam boat task? _____

Not

How many feet of water does Mark Twain mean? _____

LABELING ITEMS

These items, which are a mainstay of industrial arts and science teachers, are as nearly foolproof as any. Moreover, they are applicable not only to knowledge of tangible objects, but also to abstractions that may be presented schematically. Knowledge of governmental organization, plant and animal classification, stage directions and movements, may be tested by labeling items as well as knowledge of life cells (see Figure 10), the parts of a rifle, rivers and mountain ranges, etc.

Among the special directions to follow in designing labeling items are these.

1. The diagram to be labeled should be clearly drawn and entirely recognizable.

2. The parts to be labeled should be indicated definitely, and, as well, the place where the label is to be written.

3. Scoring often is facilitated if labels are written in a vertical column rather than on the face of the diagram. This requires the use of arrows and should not be done if the diagram will become cluttered and confusing.

4. Labeling items can be converted into select-an-answer items if desirable, by the inclusion of a list of labels (together with some extras as decoys) to be properly assigned. The item is then a special type of matching item.

ARRANGEMENT-OF-ELEMENTS ITEMS

Arrangement items (see Figure 11) have had limited use because they measure memory of relationships and organization or the ability to reason out such relationships. Too, assembly items are most appropriate to mechanical subjects in which observation and product analysis often are more efficient means of measurement. Both ordering and assembly items are wasteful of time and space, and scoring them may be difficult if partial credit is to be given for partially correct arrangements.

However, the items seem to have an unusually high potential for measuring concepts of organization in many school subjects and for measuring reasoning and spatial relations aspects of intelligence. As such, they deserve full consideration by any test designer.

In addition, arrangement type items are useful in the measurement of personality variables. The MAPS test (30), one of many that employ arrangement items, consists of many pictures from which the subject is asked to select a few and to arrange them in a picture story. The pictures he selects and the way he arranges them are taken to be indicative of his personality structure and tensions.

The following principles bear on the construction of arrangement-of-elements items.

1. The type of arrangement desired should be stated clearly and be illustrated if necessary. Illustrations usually *are* necessary for elementary pupils.
2. The place where the desired array is to be put or presented should be indicated definitely.
3. Research furnishes no precise information as to how many elements should go into an arrangement item for given purposes. However, we know generally that the number of elements appropriate for the same relative degree of difficulty should increase from grade to grade. Moreover, enough elements should be used to minimize chance solutions. If three are given, chance would produce a solution one out of six times on the average. If four are given, the chance factor decreases to one in twenty-four.
4. As a rule, an arrangement item deserves special weighting if it is to be included in a total score along with items requiring less thought and time.

Composition of the Test

Assemble items and affix directions according to the intended use of the instrument and certain tenets of test format. While valid items may have been devised that sample adequately all dimensions of the phenomenon being measured, it still remains to assemble these into a "test." The characteristics of the test as a whole are of equal importance with the characteristics of the items in determining the efficiency of the instrument. *Improper* length, se-

quence of items, directions, scoring key, etc., can invalidate the test results. On the other hand, *proper* length, sequence, directions, and scoring key will contribute validity and reliability over and above that inherent in the items.

LENGTH

Two primary considerations are involved in the length of a test: optimum administration time and reliability. From Table 2, it is apparent that certain types of item may be administered more rapidly than others. Hence, for a given administration time, a true-false test may be relatively long and a short-answer test relatively short.

Administrative Considerations. Secondary school periods vary in length but fifty minutes probably is close to the median. It follows that a test designed for a secondary grade ordinarily should be administrable in such a space of time. If we allow five minutes to begin the class, and five to close it, maximum test time for a fifty-minute period is forty minutes. If a longer test is necessary than may be taken in one period, the test may be split into self-contained halves and administered on successive days. In elementary grades, available periods of time may be longer but the attention span of children certainly is shorter. Thus, for administrative efficiency, optimum test length at the elementary level should be no longer than at the secondary and in most cases somewhat shorter.

It is necessary to consider pupil fatigue and interest in setting the length of a test as well as the time available. Unfortunately, the varied nature of pupils and of tests precludes any prescriptive statements as to fatigue and interest. The individual teacher simply must judge how long his pupils can maintain a good test set for a particular subject at a particular grade level. In general we know that older children usually can sustain a good test set for a longer period than younger children. A variety of types of test item will permit longer sustained attention than a single type. Brighter and better informed students can be tested for a longer period with good rapport than can slower or less informed.

Obviously, if all pupils are supposed to finish a test, its length should be geared to the performance of the slowest pupils. If the test is to measure speed of performance, its length should be such that the fastest pupils cannot complete it in the allotted time. The rate of the swiftest pupil may be measured only if a rate exceeding his is possible for the test.

Reliability Considerations. As for the reliability consideration in test length, select-an-answer tests achieve more reliability as they become longer simply because the chance factor in responding becomes relatively less significant. Moreover, longer tests comprised of any type of item provide a larger sample of the dimensions being measured than do shorter tests and, hence, are likely to be the more reliable. An analogy may serve to explain the importance of test length for test reliability.

Suppose that a nurse was directed to obtain a patient's "true" blood pressure. One use of a sphygmometer (the familiar tape, bulb, and meter) would

be subject to chance factors in the nurse's readings and to situation variables in the patient's blood pressure (excitement, state of digestion, fatigue, etc.). A second application would be subject to similar chance and situation variations, as would a third, fourth, and so on. But an average of two readings should be more reliable than either alone, since it is likely that the chance and situation variables would not make for errors in the same direction each time. An average of three could be trusted even more, and so on, until an average of some large number of blood pressure readings could be treated as the patient's true blood pressure: true because, over the long run, chance and situation variations would occur in opposite directions and in equal magnitude and, thus, could be assumed to cancel out. Let each separate blood pressure measurement represent one test item, the total number of blood pressure measurements represent the total number of items in the test, and the average blood pressure reading represent a pupil's score on the test. Obviously, the greater the number of items, the more reliable is the test score.

Formulas have been devised to determine the number of additional items needed to raise the reliability of a test by a given amount.¹⁹ These mathematical calculations are applicable only if the items to be added have reliability equivalent to that of items present. It is well to know that the relationship between length and reliability is one of diminishing returns. Double the length of a very unreliable test and you may have a substantial increase in reliability; but double the length of a test whose reliability already is very high and your gain is slight.

ITEM SEQUENCE

In general, the sequence of items should be such as to engender a favorable test attitude, and to provide no clues to correct or desired responses. Given types of items should be grouped together (i.e., all multiple choice, then all true-false, etc.) both for clarity of directions and for ease of pupil response. As we have stated earlier, no pattern of right answers should be detectable. As a rule, this absence of pattern may be accomplished by establishing the sequence among items of a type in some chance fashion.

If several items are to relate to a subdivision of the subject, it is debatable whether these should be grouped together or scattered throughout the test. There is a greater possibility of contextual clues to correct responses if they are grouped together, but such grouping is logical and probably makes for better test rapport.

¹⁹ One such is the Spearman-Brown prophecy formula,

$$r_d = \frac{Nr_o}{1 + (N-1)r_o}$$

in which r_d = the desired coefficient of reliability, r_o = the present coefficient of reliability, and N = the proportionate increase in length necessary to obtain the desired reliability. The meaning of a coefficient of reliability is explained on pages 184-186.

DISTRIBUTION AND RANGE OF DIFFICULTY OF ITEMS

It is important in tests of achievement that some items poor achievers are likely to know occur among the first ten or so items of the test and that easy and difficult items are distributed evenly throughout the test. The frustration to be found in reading item after item after item that he can't answer correctly may cause a poor pupil to stop trying. An even distribution of easy and difficult items throughout the test does not hold for so-called power tests. These instruments are designed to be progressively more difficult and thus to measure a pupil's performance by the last item he can answer.²⁰

Except for power and speed tests, it is thought that the items of a test should have an average difficulty of about 50 per cent. As many should have greater than 50 per cent difficulty as have less than 50 per cent difficulty, or all should have 50 per cent difficulty. The advantage in this condition is that median or mean scores (see p. 149) then will fall about halfway between the least probable score and the maximum possible score. With the median and mean at such a mid-point of the possible range, pupil scores then have the greatest freedom to assume a "natural" distribution—not to bunch up either toward high scores or low scores simply because of the nature of the test.

It is appreciated that this proposal may be contrary to practice in many classrooms. Tests often are prepared so that 70 per cent is to be the passing mark, A is to be 95–100 per cent, and so on. This means that the average item difficulty usually will be in the neighborhood of 80 per cent rather than 50 per cent. The use of an arbitrary passing mark (70 per cent or some other) is discussed in Chapter 9, page 194. There it will be seen that the practice has questionable aspects and may be inconsistent with valid evaluation as we conceive it. So, the advice remains: compose tests of items whose mean difficulty is about 50 per cent. Scores derived from such tests may be converted easily into any conventional system of letter marks.

TIMED TESTS

Timed tests may be employed to measure rate of performance—how fast a pupil can read, type, etc.—and for this employment they present no particular problems. Since speed is being measured, the pressure of time is justified. If frequency of error increases for certain pupils under this pressure, scores are still valid.

In some instances, however, rate of performance is not in question and yet timed tests are used. The reasons for such use are various: to get the test over in a short period of time, to maintain standard test conditions for several classes, to spur pupils to work more rapidly, and to prevent hesitant

²⁰ In measuring personality attributes, many other factors than these are involved in item sequence, but their consideration is beyond the scope of this text. See Gordon (16) for one interesting commentary.

pupils from persisting in fruitless deliberation over how to respond. Whatever the reason, when a stop-watch is held on pupils and they know that time will be called whether they have finished or not, some pupils are going to respond with anxiety or even panic. This emotional variable will, of course, decrease the scores of certain pupils by unpredictable amounts. Hence, unless speed of performance is being measured, it is thought that timed tests should be used sparingly. When they are administered, special attention should be given to rapport and a close watch kept for symptoms of distress. If it is apparent that a pupil's performance has suffered because of the time factor (*and rate of performance is not in question*) his test score might be disregarded and his achievement measured by another means.

DIRECTIONS

An axiom to follow in preparing test directions is that they should be clear and entirely meaningful to the least apt pupil who will take the test. If the test is to be reused, they should appear on the face of the test. If the test is for one administration only, they may be given orally and/or written on the blackboard. The best standards of clear and simple exposition should prevail in directions and, in addition, attention should be given to the following specifics.

1. Provide a place for pupils' names and other identifying data as desired, and orally instruct pupils to write their names on the tests.
2. Give directions for responding to different types of items at the time each different type is encountered.
3. If the pupils are not thoroughly familiar with a given type of item, *give one or more examples of how to answer.*
4. If the test is more than one page, direct the pupils to proceed to the next page at the bottom of the page. If pupils should wait for a signal to turn the page, place this direction at the bottom of the page.

SCORING KEY

The simplest scoring key is merely a copy of the test with right answers marked or written in. Efficiency is gained in scoring if a strip of right answers is used rather than the whole test page, or if a scoring stencil is employed. These two types of key are illustrated in Figure 15. Use of a stencil requires that responses be made by marking letters, numbers, or spaces that occupy different positions on the page. The actual key for a machine-scored test is prepared by the operator of the machine, but he must be furnished a machine answer sheet with correct options marked.

GUESSING

We have seen that items of the select-an-answer type have an appreciable chance factor, and consequently the issue of "guessing" is raised whenever tests are composed of true-false, multiple-choice items, and the like. There

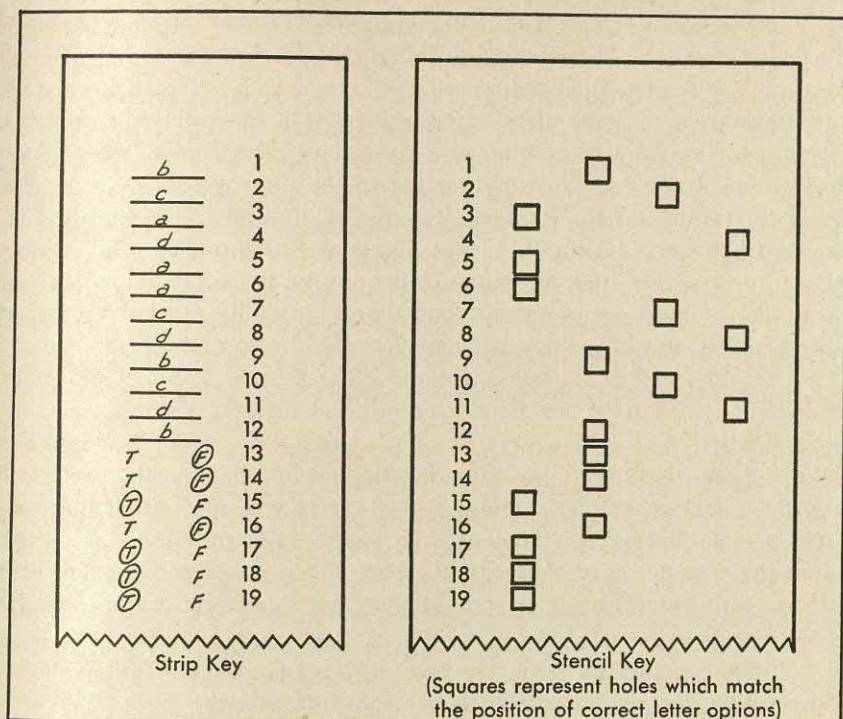


Figure 15. Example of a strip key and a stencil key.

are in existence several formulas for correction for guessing.²¹ Moreover, some current writers on educational measurement insist that guessing should be prohibited by directions and penalized if it occurs. Cook, summarizing research in achievement testing and stating conclusions based on his own research (9:1475), asserts that prohibition of guessing tends to increase the

²¹ Among the scoring formulas used to correct for guessing are the following:

For true-false items:

$$\text{Score} = \text{right} - \text{wrong}$$

$$\text{Score} = \text{total} - 2 (\text{wrong})$$

$$\text{Score} = \text{total} - 2 (\text{wrong}) - \text{omitted}$$

For four-choice multiple-choice items

$$\text{Score} = \text{right} - \frac{\text{wrong}}{3}$$

$$\text{Score} = \text{total} - \frac{2 (\text{wrong})}{3}$$

$$\text{Score} = \text{total} - \frac{2 (\text{wrong}) - \text{omitted}}{3}$$

For other number of options, and for matching items, substitute in the denominator the number of options less one. These formulas assume that wrong answers are the result of guessing. Hence, wrong responses are penalized more than omissions.

validity of a test and depresses its reliability slightly, if at all. For the prohibition to affect validity and reliability most favorably, it is essential that the time limit of the test approach an optimum figure.

It is necessary to agree that guessing may lessen the validity of a test. If guessing could actually be eliminated and, in the process, no other irrational factors were introduced, we would concur in the directions against guessing and the corrections for it. However, it is our opinion, in which several other writers may concur (21, 29, 33), that it cannot be eliminated and that correction formulas have little real value. Directions not to guess and penalties for guessing are perhaps no more important than personality attributes in determining whether a pupil will or will not guess. Moreover, the distinction between a reasoned judgment and a guess is often blurred, even for the pupil. Consequently, we advise that neither don't-guess directions nor scoring corrections be used.

When directed not to guess, it is probable that timid and submissive pupils pay more attention to the directions than aggressive and dominant ones. Moreover, good but timid pupils are likely to omit a large number of answers because they are not *exactly sure*, while poor but aggressive pupils are likely to answer a disproportionately large number that should be omitted and further depress their score. Thus, directions not to guess are likely to impose rather severe penalties on both the good timid pupil and the rash poor one.

If, on the other hand, no directions are given one way or another about guessing, personality differences are not so likely to be accentuated. And, if pupils' scores are simply the number they get correct, personality differences are not going to have a magnified effect on the score. With or without use of a correction formula, the chance factor in any score derived from select-answer items can be exactly calculated.

ADMINISTRATION OF GUIDED RESPONSE TESTS

Administering a measuring instrument to pupils requires care and skill no matter what sort of device it is. Guided response tests, however, lay the greatest claim to standardized results and, hence, are the most dependent of all on proper administration. A test essaying to measure the right dimensions and properly composed of items valid for these dimensions can still produce erratic scores through errors in administration—distractions, cribbing, class tension, and the like. Only with efficient administration can tests achieve the reliability and validity inherent in them.

Common sense and the work-a-day experience of many teachers and psychometrists are the source of rules and advice for test administration rather than experimental research. However, there is essential agreement among authorities relative to principles of good administration, and many of them are self-evident.

Equivalent Opportunity

A fundamental rule in group testing is that each pupil should have equivalent opportunity to be properly measured. None should have their scores affected negatively by administration variables nor should any be especially favored by the manner of administration.

If any directions are given orally (and some nearly always are) children who *hear* imperfectly should be seated close to the front of the room and on the side that favors their best ear. Pupils with faulty vision uncorrected by glasses need, of course, to be placed where they can best see the blackboard if any directions are to be written there. If the testing room has uneven lighting, pupils with poor eyesight should be seated where the light is best.

Sufficient Materials

It goes without saying that there should be on hand when testing begins sufficient test blanks, answer sheets, scratch paper, and whatever other materials are to be used. Since pupils will break their pencil leads or will fail to have pencils at all, a supply of sharpened pencils is a requirement. This will eliminate the need for pencil borrowing and for excursions to the pencil sharpener, all of which are unnecessary distractions.

Proper Facilities

The ordinaries of instruction seem to demand good lighting, good seating, good ventilation, etc. For the administration of a test, it is particularly essential that facilities be the best possible. Often a single test period has greater significance for a pupil's progress and handling than many instructional periods. Thirty foot-candles of diffused light on each pupil's writing surface is a widely accepted standard for the type of reading and writing usually involved in tests (19). A room temperature of 68° to 72° with some circulation of air and low humidity is likely to make pupils feel most comfortable. Seating should be spaced so that no pupil is tempted to copy from his neighbor's paper. If possible, a chair or desk space should be between each pupil and every other.

Directions

Direction for responding to a test should, as a rule, be written on test blanks or on the blackboard. When they are particularly complicated or when they are addressed to elementary grade children, it is advisable to give them orally as well. As a precaution against inattention, any oral direction should be repeated at least once, and opportunity should be given for questions before testing begins.

It goes without saying that the test administrator must be thoroughly familiar with the directions himself. For self-devised tests this is assured. With standardized tests and other imported instruments, it is well to read the test

manual completely, to read the directions on the face of the test several times and, best, to take the test yourself before you administer it to pupils.

A timed test is effective only if time limits are strictly obeyed. It is most efficient to use a timer or a stop-watch. Lacking either of these, a watch or clock with clearly marked minutes and a sweep-second hand may be employed. Do not depend on small wrist watches with "artistic" dials for tests where short intervals of time are involved. When the period involved is relatively long, there is, of course, less need for minute and second precision.

Quiet

Other things being equal, proper test administration requires quiet. Not only should pupil-generated noise be eliminated, but noises from outside the test room should be reduced to a minimum. Shouts and cries from the playground, the bustle and babble of pupils passing in the hall, the clanging of lockers, bells ringing, and the other usual noises of a modern public school all serve to disturb pupils and to reduce the validity of measurement.

Humor

A test situation, of all instructional situations, is most likely to be filled with frowns, perplexities, frustrations, fears, and many other depressing elements. These factors do not make for better test performance and they may make for poorer test performance. Consequently, a teacher should feel obliged to offset the overseriousness of many pupils and their resultant tension through smiles, a pleasant and cordial voice, courtesy, and a relaxed manner. Humor properly chosen and properly timed is a legitimate and prized ingredient in the administration of a test. This humor is thought best if it is incidental and deft. Moreover, laughter that would distract pupils from their test task should be avoided. Obviously, any jokes or witticisms should be impersonal and positively toned. Derogatory humor must be considered completely taboo during the administration of a test.

Handling Upset Pupils

Despite your best efforts, it is likely that some pupils at some times are going to become frightened, be nauseated, get dizzy, cry, and even faint during a test. Be alert to warnings of distress and act to relieve the pupil before his distress becomes acute. Flushing or loss of color in the face, hand wringing, heavy breathing, taut muscles in the neck are some of the symptoms of test malaise.

A number of simple devices are available to reduce the tension making the pupil ill. Casual assurance, either by word or gesture, that he is doing all right is all that some pupils need. It may be appropriate to ask the pupil how he is getting along and to re-explain some direction. Other pupils may benefit from your identification with their distress ("If I had to take this test, I'd be

excited too," or "When I was in school I used to turn green too but I always did better than I expected.").

In cases judged to be extreme, it may be advisable to tell the pupil that if he does poorly you will give him another chance when he feels better. It never is advisable to tell a pupil that he *isn't* frightened, that he *shouldn't* be upset, or otherwise to belittle his plight or to antagonize him. Test anxiety is bad enough without the additional irritation of a teacher's intolerance.

Should preventive measures fail and a pupil's distress be extreme, he should be allowed to quit the test and possibly to leave the room. In doing this, assurance of understanding and sympathy should be given. The pupil should be told that another opportunity to take the test will be given him or that an alternate measuring procedure will be used. To quit a test because of emotional distress necessarily is an ego-deflating event. If, in addition, a pupil feels that he will fail or miss some important academic target because of his distress, he is doubly injured.

STANDARDIZED TESTS

For perhaps fifty years now, American teachers have been able to use educational measuring instruments devised by others and claiming some degree of scientific validity. Binet and Simon issued their first intelligence scale in 1905. Earlier, Rice had used standard spelling tests in an experiment, and slightly later, after Thorndike had published a book on mental measurement, came such pioneering tests as the Courtis Arithmetic Computation Test (1909), the Ayres Handwriting Scale (1911), and the Hillegas Composition Scale (1912). The Army Alpha and Beta tests of intelligence were applied to Armed Forces draftees in 1917 and 1918, in a colossal program of mass testing, the findings and techniques of which served as data and points of controversy for years to come. The 1920's were, perhaps, the "golden years" of the standardized test. Score upon score of new titles appeared on the market; the use of standardized tests seemed to many to be a panacea for all educational ills; and, regrettably, too many tests were devised, used, and their results applied without due regard to the tenets of valid measurement.

Now, in the 1950's, users of standardized tests have a far more critical attitude toward the tests and, because of the inclusion of principles and techniques of measurement in teacher training curriculums, they are used with far greater understanding and skill. Moreover, research in test construction has provided test designers with construction and validation techniques unknown to their predecessors.

At the middle of the century, the publication of educational and psychological tests continues to increase. Oscar Buros' *Fourth Mental Measurements Yearbook*, 1953 (6) lists 793 titles, whereas the previous Yearbook

published three years earlier had listed only 663 tests. In this most complete and authoritative of all test bibliographies, the tests of 152 different publishers are reviewed. Nearly an equal number of test publishers are listed, none of whose tests happen to be described. Of the publishers, eight major firms²² seem to produce the majority of tests that enjoy widespread school use. A great many noncommercial agencies print and distribute standardized tests: college and university presses, the armed forces, state departments of education, professional organizations, and many large school districts.

Characteristics of Standardized Tests

Just the printing of a test and its general distribution does not make it a standardized test. To deserve to be called "standardized," a test must meet three critical conditions. It must have been preadministered to a population with known characteristics. It must have been revised or at least viewed critically in the light of the results of this preadministration. From the preadministration, a table of raw scores matched with correlative derived scores must be prepared and available to any user of the test. These derived scores, called norms, usually are percentile ranks or standard scores. In the case of some tests of personality attributes, they may be classifications of interest, attitude or neurotic tendency. (See Chapter 7, pages 155-160, for a full discussion of derived scores.)

"Standardized," in test construction, means about what it does in factories and laboratories. The instrument has been prepared under known and controlled conditions, the results of its use are predictable, and the significance of any measure it yields is known in advance.

Obviously, then, many of the tests listed in Buros, in other bibliographies, and in publishers' catalogues are *not* standardized tests. Almost invariably, a reputable standardized test has an accompanying manual to explain the test and to present its norms. A simple and preliminary way to determine whether a test actually has been standardized or not is to note the presence or absence of a manual.

Sources of Standardized Tests

The volume just cited, *Mental Measurements Yearbook*, is the largest and most critical comprehensive bibliography of standardized tests. All the large test publishers issue catalogues, as do many of the minor ones. Periodicals such as *Educational and Psychological Measurement* and the *Journal of Consulting Psychology* publish notices and reviews of new tests.

Specimen sets of nearly all tests are available at nominal cost. In some

²² Bureau of Educational Research, State University of Iowa, Iowa City; Educational Test Bureau, Educational Publishers Inc., Minneapolis, Minn.; Educational Testing Service, Princeton, N. J.; Houghton Mifflin Co., Boston, Mass.; Psychological Corporation, New York; Public School Publishing Co., Bloomington, Ill.; Science Research Associates, Chicago, Ill.; World Book Co., Yonkers, N. Y.

cases, the endorsement of a principal or other school official may be required before the specimen will be transmitted. It is advisable, if not mandatory, to examine a specimen before ordering a quantity.

The expense of standardized tests varies greatly, from the rather high cost of many clinical instruments (\$16 for the Binet, \$60 for the Arthur) to the 15 or 20 cents a copy for some group achievement tests. The chief expense of standardized tests is in the test booklets and manuals. Separate answer sheets may cost only 2 to 6 cents a copy.

Merits and Shortcomings of Standardized Tests

The chief value of a standardized test over a nonstandardized one lies in the added care with which it usually has been prepared, in its empirically tested reliability and validity, and in the fact that it may yield scores of general rather than purely local significance. On the debit side, a standardized test may not be addressed precisely to the dimensions you would like to measure. Its norms may be inappropriate to your pupils. And knowledge that a standardized test covering given content is to be used may have a restrictive effect on a course of study.

Now, of course, if a teacher or an official or committee in a school district can design and standardize its own test, it may be possible to achieve all the advantages of the standardized test and yet incur none of its liabilities. The standardization of a measuring instrument is a tedious and time-consuming occupation but it is an entirely possible and practical undertaking for a school or a teacher. The procedures to use are relatively simple and are fully described in many texts (2, 24, 37). Those relating to item construction and revision and to determination of reliability are presented in this volume. While more knowledge of statistics is required than might be gained from a course in measurement or even an introductory course in statistics, test standardization is not a task that requires a professional statistician.

In only one area is the use of a standardized test mandatory, this being intelligence. A measurement of intelligence is hardly worthwhile unless general significance can be claimed for it. Moreover, to devise a reasonably valid test of intelligence, let alone standardize it, requires training and experience beyond that possessed by the usual teacher, administrator, or even psychometrist. In assessments of occupational interest, of specialized aptitude (music, mechanics, etc.), and of personality patterns, it is usually advisable to employ standardized tests rather than self-devised ones. However, the available standardized tests of these phenomena are likely to be far less precise than those available for intelligence.

Judging Standardized Tests

The worth of a standardized test may be judged on the same basis as any other measuring procedure: reliability, validity, and efficiency (see pages 40-44). These may be assessed prior to first use only by reading what the

publisher has to say about his test in catalogue and manual and what the reviewers have to say. As a rule, *completeness* of information in a manual is in itself an evidence of worth. Absence of information may be construed to mean that a given factor has not been assessed (which is bad) or that an adverse finding has been withheld (which is worse).

INFORMATION OF IMPORTANCE

Consequently, as a first criterion, detailed and precise information about reliability, validity, and efficiency should be available for the standardized test in question. Bearing on these are the nature and size of the standardization population, the techniques employed in item construction and revision, and the dimensions that the test purports to measure; so these too should be fully described.

RELIABILITY

The reliability of a standardized test may be stated mathematically as a coefficient of reliability (see page 185), but the degree of reliability necessary depends upon the use of the test and the reliability of like tests. For appraisals of groups of pupils and decisions about groups, lower reliabilities are acceptable. For measurements of individuals and important decisions affecting them, higher reliabilities are desired. Many intelligence and achievement tests have coefficients of reliability of .90 to .95, but the reported reliabilities of personality tests tend to run in the .80's.

VALIDITY

Validity sometimes *is* stated mathematically, as a derivative from a coefficient of reliability and/or as a coefficient of correlation between the test and some criterion. If a suitable criterion exists, a validity coefficient relative to it has some meaning, but oftentimes the criterion has no more pretense to validity than the test with which it is compared. For intelligence tests, comparisons between the given test and a battery of other tests and/or the Wechsler or Binet are important for validity. Achievement tests should be expected to correlate to a high degree with a composite of many other reputable achievement tests covering the same subjects. Moreover, they should show a high positive correlation with school marks in the subjects in question. In the writers' view, validity coefficients derived from reliability coefficients still are reliability coefficients and deserve no separate attention.

Very often, no mathematical expression can be given to a test's validity, and the case for it rests upon a discussion of its construction and the data resulting from its administration. In weighing this discussion, attention should be given to any analysis of item discrimination and difficulty, to the relationship between test scores and age and school grade, and to the shape of raw score distributions (see pages 98-101, 186). Even if a coefficient of validity is presented, a description of these points is necessary.

A test's validity is not just an inherent factor; it relates necessarily to the use to which the test is to be put, and to the similarity between the population used for standardization and the pupils to whom the test now is to be applied. Obviously, the test will have validity for *your purpose* somewhat to the degree that there is agreement between the dimensions that the test purports to measure and the dimensions you wish to measure. And the raw scores of your pupils may be converted into valid derived scores using the test's norms only to the degree that your pupils possess the same essential characteristics as the standardization population upon which the norms are based. Conversion of a raw score on a standardized test into a percentile or a standard score from the test's table of norms simply gives the pupil a position in the standardization group. This position is meaningful only when the pupil is within the same age range, has had about the same experience, possesses much the same motivation, etc., as the pupils upon whom the norms were based. If these similarities are missing, we may not be sure *what* caused a pupil's score and we may *not*, with safety, go beyond the simple fact that he made a relatively high, low, or intermediate raw score on the test.

The efficiency of a standardized test involves its administration time, the ease with which it may be administered, how it may be scored, and its cost. Such factors usually are described in manuals and even in catalogue entries. To judge the efficiency of a given test for your use requires merely that these factors be weighed against the available funds, time and facilities.

Some Special Considerations in Using Standardized Tests

Many tests that cover several dimensions or subject aspects (such as spatial perception, reasoning, etc., in intelligence; vocabulary, rate, and comprehension in reading) provide for the recording of part scores and for their representation on a graph. These score graphs are called profiles and one is illustrated in Figure 16. Profiles permit easy visualization of the variations in

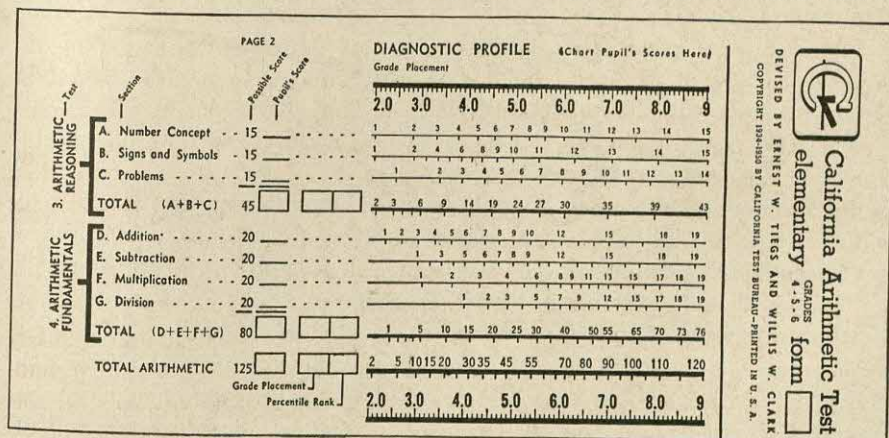


Figure 16. A test profile form. (Copyright 1934-1950 by California Test Bureau.)

a pupil's achievement, interests, or intellectual facilities, and they facilitate the diagnostic use of measurements.

As all elementary teachers know, and as all of us should remember, there are standardized tests that measure achievement in several school subjects. Buros lists twenty-six such "Achievement Batteries." Nearly all of them cover reading, vocabulary, and arithmetic. To these, some add science, social studies, English usage, and spelling. As a rule, an elementary achievement battery is produced in several forms for use at different grade levels. If achievement batteries are to be used for measurement at several points in the pupil's progress through school, it is advisable that appropriate forms of the same basic test be used rather than different tests. This permits more consistent charting of progress and affords more valid comparisons between scores made at different grade levels.

Summary

"Guided response procedures" is used as a categorical term for true-false, multiple-choice tests and the like. The essence of such procedures is that testees have only a limited number of options for response to any item and the significance of each option is predetermined. There seem to be three basic types of guided response item: select-an-answer (true-false, matching, multiple-choice, etc.), provide-an-answer (fill-in, short answer, labeling, etc.) and arrangement of elements (ordering, assembly, etc.). The procedures make use of all the essential devices of behavioral measurement. The questions themselves are *standard stimuli*. Limited options of response with predetermined significance constitute *standard responses*. The items may be scored with a key and thus they embody *standard analysis*. Sampling, of course, occurs because items in a test are less numerous than the components of achievement, personality, etc., toward which they are directed.

The procedures yield test scores that may be considered as classificatory numbers and usually can be converted into numbers indicative of rank. Scale symbols are possible under certain special circumstances.

Construction of valid and reliable guided response tests necessitates adherence to some such directions as these.

1. Define exactly the phenomenon to be measured and its measurable dimensions in behavioral terms.

2. Devise verbal and/or graphic items, the responses to which automatically will classify pupils with respect to given dimensions. The classification usually is dichotomous and in achievement tests the two categories usually are "knows it" and "does not know it." To accomplish this classificatory function accurately, guided response items should be short, affirmative, simple, and unequivocal; couched in pupil language; *not* lifted from text books. Collectively, the items should discriminate among all types of individuals to be tested; be mutually independent; sample adequately all dimensions of the phenomenon being measured that are subject to sampling.

3. Select items for use according to their special properties and construct them according to their type specifications. The types vary as to chance of guessing, as to time of administration, and as to difficulty.

4. Assemble items and affix directions according to the intended use of the instrument and certain principles of test format, length, item sequence, timing, directions, scoring key, and the operation of chance.

The administration of guided response tests requires care and skill if the condition of standard stimulation and standard response is to be attained. Each pupil or other testee should have equal opportunity to be measured properly. Sufficient materials should be at hand. Lighting, seating, ventilation, etc., should be optimum. Directions usually should be given orally as well as in writing. A quiet and relaxed atmosphere should prevail.

"Standardized test" is the collective name for guided response instruments that have been validated through trial and for which norms have been developed. The norms are the scores achieved by a population having known characteristics or statistics derived from such scores, such as age means, grade means, percentile equivalents, etc. The chief advantages of standardized tests over locally devised ones lie in their technical excellence, their "national" norms, and their usually high reliability. On the weakness side is their too frequent inappropriateness to the objectives of a school, the organization of its curriculum, or the character of its pupils.

EXERCISES

1. For a subject in which you specialize, prepare several test items of each of the following types:

- True-false
- Multiple-choice
- Matching
- Completion or fill-in
- Short answer
- Labeling
- Ordering or assembly of elements

2. Prepare one or more guided response questions that will measure reasoning about facts and not just memory of facts.

3. Devise a set of guided response items that combines two or more of the basic types listed in Exercise 1.

4. Study the manual of a standardized test, administer it to a group of pupils, and score the papers.

5. Define some aspect of pupil achievement you would like to measure, state its dimensions in behavioral terms, and construct a guided response test that will measure pupils relative to these dimensions. Indicate the dimension to which each test item is keyed.

6. Criticize a guided response test that another person has prepared in terms of the principles for test construction outlined in this chapter.

BIBLIOGRAPHY

1. Anderson, S. B., "Sequence in Multiple Choice Item Options," *Journal of Educational Psychology*, 43:364-380, 1952.
2. Bean, Kenneth L., *Construction of Educational and Personnel Tests*. New York: McGraw-Hill Book Co., Inc., 1953.
3. Bender, W., and Davis, R. A., "What High School Students Think About Teacher Made Examinations," *Journal of Educational Research*, 43:58-65, 1949.
4. Bloom, B. S., and Heyns, I. De V., "Development and Applications of Tests of Educational Achievement," *Review of Educational Research*, 26:71-88, February, 1956.
5. Brinkmeier, I. H., and Ruch, G. M., "Minor Studies on Objective Examination Methods: III Specific Determiners in True-False Statements," *Journal of Educational Research*, 22:110-118, 1930.
6. Buros, Oscar K., *Fourth Mental Measurements Year Book*. Highland Park, N. J.: Gryphon Press, 1953.
7. Cattell, R. B., "Objective Measurement of Dynamic Traits," *Educational and Psychological Measurement*, 10:224-248, 1950.
8. Clark, E. L., "General Response Pattern to Five-choice Items," *Journal of Educational Psychology*, 47:110-117, February, 1956.
9. Cook, W. W., "Achievement Tests," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 1461-1478.
10. Coombs, C. H., "Concepts of Reliability and Homogeneity," *Educational and Psychological Measurement*, 10:43-56, 1950.
11. ———, "On the Use of Objective Examinations," *Educational and Psychological Measurement*, 13:308-310, 1953.
12. Cronbach, Lee J., "Further Evidence on Response Sets and Test Design," *Educational and Psychological Measurement*, 10:3-31, 1950.
13. Ebel, R. L., "Obtaining and Reporting Evidence on Content Validity," *Educational and Psychological Measurement*, 16:269-282, 1956.
14. Flanagan, J. C., "Use of Comprehensive Rationales in Test Development, Educational and Psychological Development," *Educational and Psychological Measurement*, 11:151-155, 1951.
15. Gleser, G. C., "Speed of Response as a Measure of Difficulty," *American Journal of Psychology*, 64:548-563, 1951.
16. Gordon, L. V., "Effect of Position on the Preference Value of Personality Items," *Educational and Psychological Measurement*, 12:669-676, 1952.
17. Hale, William H., "The Quizmasters Fasten Onto Higher Education," *The Reporter*, October 6, 1955, pp. 14-20.
18. Hawkes, H. E., Lindquist, E. F., and Mann, C. R., *The Construction and Use of Achievement Examinations*, Boston: Houghton Mifflin Co., 1936.
19. Holy, T. C., and Herrick, John H., "School Plant," in *Encyclopedia of Educa-*

- tional Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, p. 1107.
20. Hurlburt, Delpha, "The Relative Value of Recall and Recognition Techniques for Measuring Precise Knowledge of Word Meaning—Nouns, Verbs, Adjectives," *Journal of Educational Research*, 47:561–576, April, 1954.
 21. Jackson, R. A., "Guessing and Test Performance," *Educational and Psychological Measurement*, 15:74–79, 1955.
 22. Kropp, Russell, "Relationship Between Process and Correct Item-Response," *Journal of Educational Research*, 49:385–388, January, 1956.
 23. Lennon, Roger T., "Assumptions Underlying the Use of Content Validity," *Educational and Psychological Measurement*, 16:294–303, 1956.
 24. Lindquist, E. F. (ed.), *Educational Measurement*. Washington, D.C.: ACE, 1951.
 25. Loevinger, J., "Systematic Approach to the Construction and Evaluation of Tests of Ability," *Psychological Monograph*, V 61, No. 4, American Psychological Association, 1947.
 26. Lorge, I., and Kruglov, L. P., "Improvement of Estimates of Test Difficulty," *Educational and Psychological Measurement*, 12:554–561, 1952.
 27. Odom, C. L., and Miles, R. W., "Oral Versus Visual Presentation of True False Items in the First Course in Psychology," *Educational and Psychological Measurement*, 11:470–477, 1951.
 28. Scates, D. E., "Changing Role of Paper and Pencil Tests in Evaluating Educational Outcomes," *High School Journal*, 34:109–113, 1951.
 29. Sherriffs, A. C., and Boomer, D. S., "Who Is Penalized by the Penalty for Guessing?" *Journal of Educational Psychology*, 45:81–90, February, 1954.
 30. Shneidman, Edwin S., *Make a Picture Story*. New York: Psychological Corporation, 1947.
 31. Sims, V. M., "Questioning Some Assumptions Underlying Current Achievement Testing," *Educational and Psychological Measurement*, 8:565–573, 1948; and *Educational Digest*, 14:29–31, March, 1949.
 32. Snodgrass, Florence T., "Unreliability of Group Test Profiles," *Journal of Educational Psychology*, 45:129–142, March, 1954.
 33. Stanley, J. C., "Psychological Correction for Chance," *Journal of Experimental Education*, 22:297–298, March, 1954.
 34. Suchman, E. A., "Logic of Scale Construction," *Educational and Psychological Measurement*, 10:79–93, 1950.
 35. Tate, Merle W., *Statistics in Education*. New York: The Macmillan Company, 1955, pp. 359–370.
 36. Tiegs, E. W., *Tests and Measurements in the Improvement of Learning*. Boston: Houghton Mifflin Co., 1939, pp. 93–95.
 37. Travers, Robert M. W., *Educational Measurement*. New York: The Macmillan Company, 1955.

CHAPTER 7

STATISTICAL DESCRIPTIONS OF MEASUREMENT DATA

We have seen in previous chapters that measurement entails the assignment of numbers or other limited symbols to the dimensions of phenomena, and we have discussed the many procedures (observation, testing, etc.) by which this assignment may be accomplished for behavioral phenomena. Now, let us examine certain mathematical operations that may be necessary before our assigned symbols are fully meaningful.

To begin, let us think of a teacher who has just rated a group of pupils as to the effectiveness of their study habits and at the same time has just scored a multiple-choice geography test completed by the same pupils. He has before him the usual first outcomes of behavioral measurement, an unsystematic array of symbols and numbers or, as we describe them technically, *raw data*. So far as the ratings of study habits are concerned, he may, if he wishes, leave the data raw. In the case of the geography test, however, the "raw" data need to be arranged and treated mathematically before he records them. The difference lies in the fact that the ratings in themselves describe the status of the pupils and thus are measures, whereas the test scores do not in themselves describe anything. They are just numbers. Before a measure is accomplished, before a pupil's classification or rank or scale position is expressed, something further must be done with them.

Just what is to be done with these raw test scores is determined by the forms of measurement we seek, by the phenomena we measure, and by the measuring procedures we have used. But in general, we manipulate the numbers so as to describe the group in terms of the individuals comprising it and the individuals in terms of the group to which they belong. We shall restrict ourselves in this chapter to definitions and explanations of basic statistical terms and processes that may be used to describe groups and individuals.

Tabular and Graphical Portrayal of Group Measurements

The simplest step for organizing numerical data in tabular or graphical form is to arrange the numbers in decreasing or increasing order of size. This is commonly called a *distribution of scores*. Typically, the highest score is written first, followed by the next highest, and so on until the lowest score is

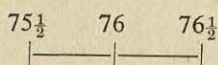
reached. If a certain score occurs more than once in the data, then it is repeated the necessary number of times in its proper position in the sequence. The following is an example of such an arrangement.

55	43	39	21
53	41	35	19
52	40	36	19
49	40	31	17
49	40	31	16

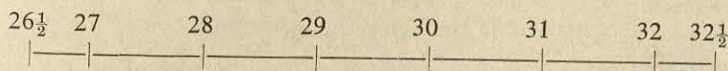
FREQUENCY TABLES

Ordinarily, for relatively small groups, such a simple arrangement would be all that is necessary for descriptive purposes or for any further calculations. For larger groups of data, say one hundred scores or more, such an arrangement would be very little improvement in way of further organization. The data would still be largely unmanageable for further calculations and whatever group pattern might exist in the data would still be unidentifiable. Distributions containing a large number of scores need a more compact tabular arrangement. This is provided for by the device of grouping the scores into *intervals* and then indicating the number or frequency of scores in each interval. This is called a *frequency table*.

For our purposes, we shall define an *interval* as any indicated grouping of numerical scores. An interval, then, may contain only one score or may contain several scores. Conventionally, if an interval contains only one score, say 76, then the interval extends from $75\frac{1}{2}$ to $76\frac{1}{2}$. Graphically, this interval would be indicated as follows:



When several scores are included in an interval, the interval is designated by a lower and an upper boundary score. For example, the interval 27–32 includes the following six scores: 27, 28, 29, 30, 31, 32; and graphically this interval would look like this:



Note that the interval extends a half unit below ($26\frac{1}{2}$) and a half unit above ($32\frac{1}{2}$) the designated boundaries. These are called the *actual boundaries* of an interval.

The *interval size* is the number of scores contained in the interval or the difference between the actual boundaries of interval, both being numerically the same. Thus, the size of the interval 27–32 is 6, because it contains 6 scores and because the difference between its actual boundaries, $26\frac{1}{2}$ and $32\frac{1}{2}$, is 6. The *mid-point* of an interval is found by taking half of the interval

size and adding it to the lower actual boundary or subtracting it from the upper actual boundary. For example, the mid-point of the interval 27-32 is found by taking half of the interval size ($\frac{1}{2}$ of $6 = 3$) and adding it to the lower actual boundary ($26\frac{1}{2} + 3 = 29\frac{1}{2}$) or subtracting it from the upper actual boundary ($32\frac{1}{2} - 3 = 29\frac{1}{2}$). In both cases the mid-point is $29\frac{1}{2}$.

As a summary, here are some further examples that will illustrate the different aspects of intervals just discussed.

<i>Designated Interval</i>	<i>Actual Boundary Limits</i>	<i>Size</i>	<i>Consecutive Scores Included</i>	<i>Mid-point</i>
32-37	$31\frac{1}{2}$ to $37\frac{1}{2}$	6	32, 33, 34, 35, 36, 37	$34\frac{1}{2}$
52-53	$51\frac{1}{2}$ to $53\frac{1}{2}$	2	52, 53	$52\frac{1}{2}$
68-71	$67\frac{1}{2}$ to $71\frac{1}{2}$	4	68, 69, 70, 71	$69\frac{1}{2}$
40-44	$39\frac{1}{2}$ to $44\frac{1}{2}$	5	40, 41, 42, 43, 44	42

A sequence of intervals of the same size provides a convenient device for tabulating the raw scores. For instance, if the raw scores ranged from 50 to 90, then the following sequence of intervals could be used:

90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54

Note that the sequence of intervals is arranged so that there is no overlapping of scores and hence no ambiguity as to the interval to which a score belongs.

The number of intervals used in a sequence to cover a given range of scores may vary according to the purpose one has in mind for grouping the scores, and to experience and judgment. Most frequently, scores are grouped to determine the over-all pattern in the distribution. If this is the case, the following is a suggested guide for determining the number of intervals to use in a sequence.

<i>Number of Scores in Distribution</i>	<i>Suggested Number of Intervals to Use</i>
30	6 to 8
50	7 to 9
100	8 to 10
300	9 to 12
500	10 to 14
1,000	12 to 17

Once the number of intervals has been determined, it is a simple matter to compute the size of the intervals. First find the range of the scores by subtracting the lowest score from the highest. Then divide the range by the number of intervals to be used and the result is the approximate size of the interval. After the sequence of intervals has been set up, the scores are tallied in the intervals and the result is a *frequency table*.

To review the steps just described, let us begin with some "raw data" and outline the things necessary to set up a frequency table that will best describe the distribution of the scores.

TABLE 3
An Outline for Constructing a Frequency Table

Steps to follow:		Raw data		
1. Note the number of scores in the raw data. In this example there are 90 scores		119, 112, 98, 123, 100, 103, 112, 108, 105, 105		
		94, 106, 100, 109, 97, 115, 102, 107, 99, 121		
		111, 101, 108, 102, 96, 95, 110, 113, 107, 103		
		127, 104, 113, 104, 117, 106, 93, 91, 101, 104		
2. Determine the desirable number of intervals to use. Here, say 9 or 10 intervals		118, 97, 116, 101, 111, 105, 124, 109, 98, 111		
		92, 110, 108, 103, 107, 115, 92, 106, 100, 126		
		109, 101, 118, 111, 102, 97, 104, 95, 122, 119		
		115, 113, 128, 104, 133, 121, 106, 113, 110, 96		
3. Determine the range by subtracting the lowest score from the highest score		95, 107, 109, 105, 120, 103, 107, 114, 108, 116		
133 - 91 = 42				
		Frequency table		
		Interval	Tally	Frequency
		130-134	1	1
		125-129	111	3
		120-124	11111 1	6
		115-119	11111 11111	10
		110-114	11111 11111 1111	14
		105-109	11111 11111 11111 11111 1	21
		100-104	11111 11111 11111 1111	19
		95-99	11111 11111 1	11
		90-94	11111	5
4. Determine the size of the intervals by dividing the range by the number of intervals desired. $42 \div 9 = 5$ approximately				—
5. Set up the sequence of intervals				
6. Tally the scores in the proper intervals			Total	90

Now let us take a look at two other frequency tables of the same raw data. In Table A, the size of the interval is smaller and hence the number of intervals to cover the range is greater. In Table B, the size of the intervals is larger and the number of intervals required are fewer.

(A)		(B)	
<i>Interval</i>	<i>Frequency</i>	<i>Interval</i>	<i>Frequency</i>
132-133	1	130-139	1
130-131	0	120-129	9
128-129	1	110-119	24
126-127	2	100-109	40
124-125	1	90- 99	16
122-123	2		—
120-121	3	Total	90
118-119	4		
116-117	3		
114-115	4		
112-113	6		
110-111	7		
108-109	8		
106-107	9		
104-105	9		
102-103	7		
100-101	7		
98- 99	3		
96- 97	5		
94- 95	4		
92- 93	3		
90- 91	1		
	—		
Total	90		

Note that in Table *A*, the distribution is so spread out that it is rather difficult to determine any possible group pattern. Likewise in Table *B*, the distribution is so bunched up that it is still difficult to determine any group pattern. Thus we see that different numbers of intervals affect the pattern of the distribution. Somewhere between these two extreme examples lies the optimum number of intervals that will best exhibit whatever group pattern may exist in the distribution of scores. The original frequency table exhibited in Table 3 does show a group pattern or shape in the distribution. This frequency table indicates fewer scores at the upper end of the scale with most of the scores grouped at the lower end of the scale.

With just the frequency table, it is difficult to visualize the distribution of the scores. Therefore, we turn to graphic methods of portraying the data. There are many different ways of graphically portraying a frequency distribution. However, certain of these ways are particularly straightforward and effective, and will be especially useful in the later developments of further statistical measures. Among these graphical representations we shall discuss the histogram, the frequency polygon, and the smoothed frequency curve, all of which are based on the frequency table.

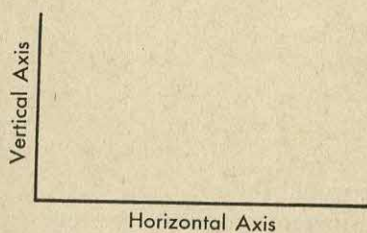
HISTOGRAMS

The *histogram* is essentially a vertical bar graph and is particularly adaptable to the intervals of a frequency table. The first step in constructing a histogram from a frequency table is to set up a horizontal axis and a vertical axis. Then on the horizontal axis, lay off the boundary marks of the intervals and on the vertical axis, lay off a scale of frequency per interval. The final step consists of erecting a vertical bar at each interval representing the frequency for that interval. An illustration will perhaps serve to clarify the steps that are involved in setting up a histogram. This is presented in outline form in Table 4.

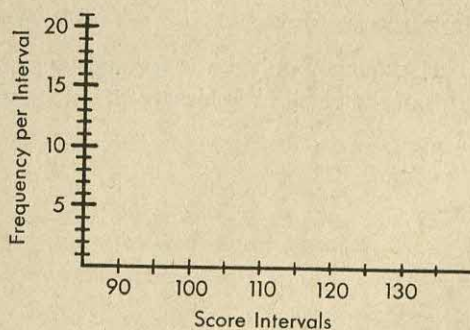
TABLE 4
An Outline for Constructing a Histogram

Steps to follow:	Interval	Frequency
1. Start with the frequency distribution you wish to represent graphically as a histogram.	130-134	1
	125-129	3
	120-124	6
	115-119	10
	110-114	14
	105-109	21
	100-104	19
	95- 99	11
	90- 94	5
	Total	90

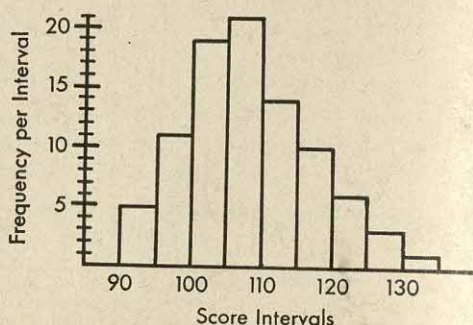
2. Set up a vertical axis and a horizontal axis.



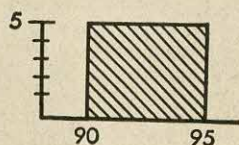
3. Lay off boundary marks of intervals on horizontal axis and frequency per interval on vertical axis.



4. Erect vertical bars at each interval indicating the frequency in that interval.



Looking at the finished histogram, we can readily see that it gives an effective visual presentation of group pattern, and is simple to construct. There are some important features to note concerning the histogram. First of all, the boundary marks on the horizontal axis are only approximate. For instance, the boundary mark 95 on the histogram should technically be $94\frac{1}{2}$. The actual boundary marks have all been rounded off to the next highest whole number, and this is done solely for convenience of representation. It is also important to note that the scale on the vertical axis represents frequency per indicated interval rather than just frequency. This necessitates the intervals along the horizontal axis being of equal size. The frequency for each interval is actually represented by the area of the vertical bar erected at that interval. As an example, for the first interval of the histogram, its frequency of 5 is represented by the shaded area of the vertical bar shown with the following portion of the histogram.



It would follow, then, that the sum of the areas of the vertical bars of the histogram represents the total frequency of 90 scores.

FREQUENCY POLYGONS

The frequency polygon is formed by simply connecting the mid-points of the tops of the bars of the histogram with straight lines, as shown in Figure 17.

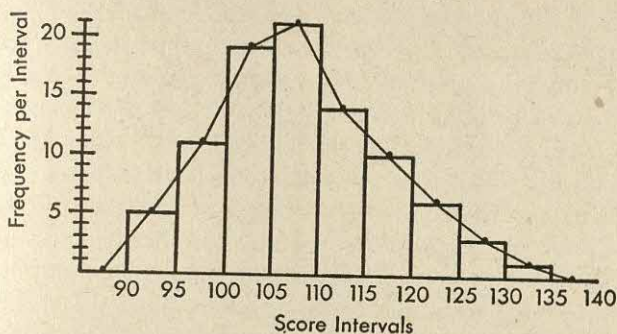
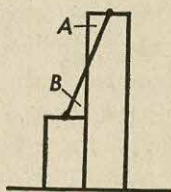


Figure 17. Frequency polygon.

By drawing the frequency polygon through the mid-points of the vertical bars of the histogram, the area under the polygon is equivalent to the area of the histogram. This is shown by the following figure in which a straight line connects the mid-points of two adjacent vertical bars.



Triangle *A*, cut off from the right bar, is equal to triangle *B*, added to the left bar of the histogram. This occurs throughout the histogram, leaving the area unchanged.

Since the frequency polygon is made up of straight lines connecting successive points, the result generally is a jagged effect. This can be offset somewhat by an increase in the number of scores. However, if this is not possible, the jagged effect of the frequency polygon may be reduced by simply sketching

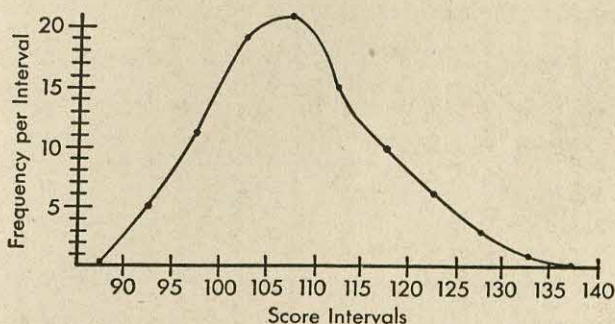


Figure 18. Smoothed frequency curve.

a smooth curve that approximates the general shape of the frequency polygon. The result is called a smoothed frequency curve and is illustrated in Figure 18. The smoothed curve is more pleasing to the eye and yet is very effective in portraying the over-all pattern of a distribution of scores. Once again the area under the curve represents the total number of scores in the distribution and the scale on the vertical axis represents the frequency for the intervals specified in the original frequency table.

The histogram, frequency polygon, and smoothed frequency curve are some of the many ways in which group measures may be graphically presented and are the ones most commonly encountered in educational measurement. At this point it might be well to exhibit and identify some of the possible patterns of group distributions.

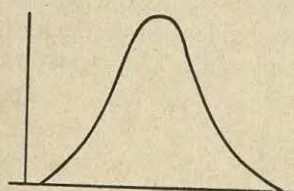


Figure 19. Bell-shaped or normal distribution.

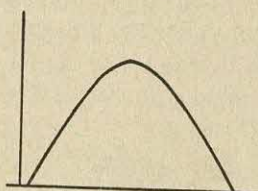


Figure 20. Symmetrical distribution.

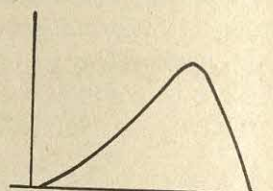


Figure 21. Skewed distribution.

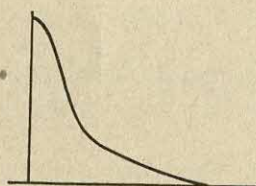


Figure 22. J-shaped or Poisson distribution.

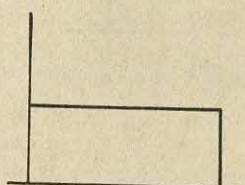


Figure 23. Rectangular distribution.

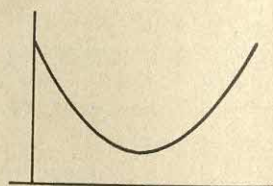


Figure 24. U-shaped distribution.

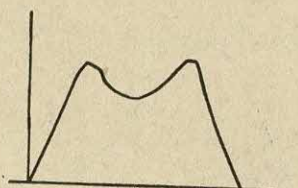


Figure 25. Bi-modal distribution.

The first (Figure 19) is a theoretical curve and will be discussed in more detail later (pages 163-169). It is essentially a curve of normal probability. The remaining figures are presented as some possible types of distributions of group measures and are identified by name.

We stated at the beginning of this chapter that the purpose of statistical presentations was to make unmeaningful measurement data meaningful, or to give additional meaning to already meaningful data. It is well to ask then, what do these frequency tables, these histograms, and these frequency polygons tell us about the pupils who made the scores that comprise the tables and graphs? What precise appraisal of status do we have that we did not have before we made the tables or constructed the graphs?

Well, from the frequency table and from the graphs we may observe which scores were lowest and which were highest. At a glance we may judge the total range of scores and the range of scores that constitute the central cluster, or we may observe that there is no central cluster. Simply by looking, we may see that the majority of the group scores are high, low, or in the middle. Finally, knowing John's or Harry's score, we can estimate his *rank* in the group and we can express some qualifications about that rank. He is at the top but it is a close distribution. He is about at the middle, but the middle is skewed toward the high end of the scores.

Describing a Group by a Representative Score

We turn now from describing groups by tables and graphs to describing groups by representative scores. The use of a representative measure is a common everyday occurrence. The average golf score of a person, say 76, is representative of that person's golfing performance. If a certain city publicizes that its mean daily temperature is 83° F., this likewise represents that city's climate. And one way for a teacher to describe the performance of her class on an arithmetic achievement test is to state a single score that best indicates the over-all performance of the group. Because it is to represent, we want that score, if possible, to be the one about which all the rest of the scores tend to cluster. For this reason, statisticians call these representative scores "measures of central tendency." In our discussion of representative scores or measures of central tendency, we shall consider the mode, the median, and the mean, which is often called the "average."

MODE

One way of selecting a representative score from a table or distribution of scores is to choose that score which occurs most frequently. When this is done, the resulting score is called the *mode*, just as the style of clothing most frequently seen is called the mode. Among the raw scores presented in Table 3, the score 105 occurs the most frequently and is therefore the mode of the distribution. When the raw scores are grouped by intervals in a table, the mode is arbitrarily defined as the mid-point of the interval in which the largest frequency occurs. Again referring to Table 3 and this time to the frequency table developed there, we find that the interval 105–109 has the largest frequency and therefore its mid-point 107 is the mode. The discrepancy between the mode of 105 determined from the raw ungrouped scores and the mode of 107 determined from the grouped scores can be explained

by the fact that, in frequency tables, the raw scores lose their individual identity and the mid-points of intervals are considered representative of all the scores in the intervals. As we progress, we shall note further discrepancies between computations based on ungrouped data and computations based on grouped data.

The mode, as defined here, is certainly simple and easy to determine by inspection. However, at best the mode is only a crude approximation of a representative score around which the rest of the scores tend to cluster.

MEDIAN

A second type of score may be used to represent a distribution of scores, that score which divides the distribution in half when the scores are arranged according to size. A score which does this is called the *median* of the distribution.

This definition of the median can be applied directly to a distribution of ungrouped scores. The first step is to arrange the raw scores in order of size. Then choose that score in the distribution above which and below which there are an equal number of scores. The following two examples will serve to illustrate the determination of the median from a few scores already arranged according to size.

	(a)		(b)	
	35		52	
	29		48	
	27		39	
The median is 20	→ 20	33	←	The median is 36
	19		27	
	11		20	
	5			

In the first distribution the median is 20 because there are three scores above and three scores below. In the second distribution, however, there is no single middle score which divides the distribution; therefore, the median is determined to be midway between the two central scores of the distribution. The two central scores in the second distribution are 33 and 39 and the midway point between these two scores is 36, which is the median of the distribution. In general, when there is an odd number of measures, then there is a single measure that divides the distribution in half. For an even number of measures, it is necessary to find the two central scores and the median is the midway point between these two scores.

If we are concerned with a large number of scores, the method just outlined for finding the median among ungrouped data would become very tedious. This will be apparent if you consider the task of arranging 1,000 scores in order of size. Hence, we need to have some method for determining the median of scores that have been grouped in intervals.

To understand the median of grouped numbers, let us set up the definition of the median in terms of a histogram. We have already observed that the area of the histogram represents the total number of scores in the distribution. Therefore, to be consistent with the previous definition, the median is now defined as the point on the horizontal axis of the histogram where 50 per cent of the area lies to the left and 50 per cent lies to the right of the point. This is illustrated in the following figure.

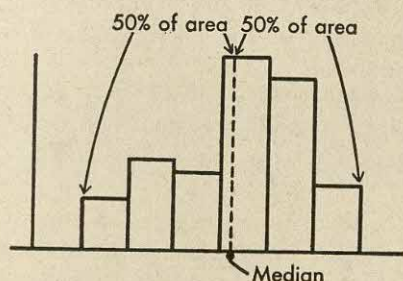


Figure 26. Median defined in terms of the area of a histogram.

Looking at the histogram in Figure 26, it is apparent that finding the median of grouped scores is essentially a matter of first determining the middle interval that contains the median and secondly of having a precise way of fixing the point within the middle interval. To see just how to obtain a median for grouped scores, let us go step by step through the process.

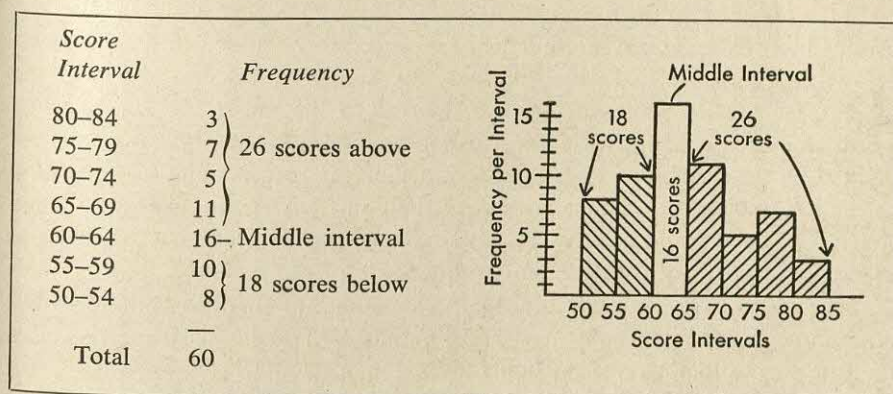


Figure 27. Finding a median illustrated.

With reference to Figure 27, our first problem is to find the middle interval that contains the median. This is done by first dividing the total number of scores by 2, in this example $60 \div 2 = 30$. There should be 30 scores above and 30 scores below the median. Now we start with the lowest interval and begin adding in the frequency column until we come to the interval where the

total frequency exceeds 30, or half the total. Adding the frequencies of the first two lower intervals we get $8 + 10 = 18$. Adding the frequency of the third interval to the frequencies of the first two, we get $16 + 18 = 34$, which exceeds 30. Therefore, the third interval from the bottom, 60–64, is the middle interval and contains the median. We can verify this by starting from the top and adding down the frequency column. Adding the frequencies of the first 4 intervals from the top, we obtain $3 + 7 + 5 + 11 = 26$. To add the frequency of the fifth interval to this sum causes the total to exceed 30. Thus again the interval 60–64 is the middle interval and it has 18 scores below it and 26 scores above it.

With the middle interval designated, the next procedure is to locate the median in the interval. We may analyze this step by using Figure 28, which reproduces the middle interval of the histogram in Figure 27.

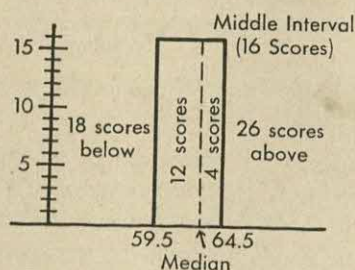


Figure 28. Locating the median in the middle interval.

From Figure 28, you can see that 12 scores need to be added to the 18 scores below the interval to make the necessary 30. Similarly, 4 scores are necessary to increase the 26 scores above the interval to the necessary 30. In other words, the area of the middle vertical bar representing the 16 scores must be divided by the median so that a portion of the area representing 12 scores lies to the left and the remaining portion of the area representing 4 scores lies to the right. This would satisfy the requirement that 50 per cent of the area of the histogram or 30 scores lie on either side of the median.

The stage is set now for the final calculation. First, note that the actual boundaries for the interval 60–64 are 59.5 and 64.5 as indicated in Figure 28 and the size of the interval is 5 units along the horizontal axis. Since 12 of the 16 scores must lie beyond the lower boundary 59.5, then the median must be located $12/16$ ths of five units or $3\frac{3}{4}$ units above 59.5. This makes the median equal to $59.5 + 3.75 = 63.25$. This may be verified by noting that 4 of the 16 scores or $4/16$ ths of 5 units must lie below the upper boundary 64.5. Thus the median is again equal to $64.5 - 1.25 = 63.25$.

Table 5 summarizes the previous detailed discussion of the steps in computing the median of a distribution of scores.

TABLE 5

An Outline for Computing the Median of a Distribution

Steps to follow:			
1. Divide the total number of scores by 2. $138 \div 2 = 69$. Thus 69 scores must lie above and below the median.		72-74	2
		69-71	13
	lower	66-68	21
	boundary	63-65	44 ← middle interval
2. Locate the middle interval containing the median by counting up or down the frequency column until the 69th score is reached. The middle interval is 63-65 and 58 scores lie below it.	62.5 →	60-62	28
		57-59	19 ↓ 58 scores lie below
		54-56	7
		51-53	3
		48-50	1
3. Find the distance from the lower boundary of the middle interval to the median. Since 58 scores lie below the middle interval then 69 minus 58 or 11 more scores of the 44 scores of middle interval are needed to reach the median. The size of each interval is 3 units. The distance is equal to 11/44ths of three units or .75 units.	Total	138	
4. Add this distance to the lower boundary of the middle interval and the result is the median of the distribution. The lower boundary of the middle interval is 62.5. Thus $62.5 + .75 = 63.25$ is the median.			

The median found in Table 5 may be verified by noting that 36 scores lie above the middle interval and therefore $69 - 36$ or 33 more scores of the 44 in the middle interval are needed. The distance from the upper boundary of the interval down to the median is 33/44ths of 3 units or 2.25 units. Subtracting this distance from the upper boundary, we get $65.5 - 2.25$ equals 63.25, which is the same as we found by starting our count from the bottom.

From the computation in Table 5 we may observe that it is possible to compute the median directly from the frequency table without referring to the histogram. As an introduction, however, the use of the histogram gives us a better visual picture of what is being done in the computation of the median. It is important to note that in the computation of the median, the scores in the middle interval are assumed to be evenly distributed throughout the interval. This assumption is certainly necessary to facilitate computation, and this is the explanation of any discrepancy between the median computed directly from the raw scores and the median computed from a frequency table of the same scores. In any case, whether computing from raw scores or

grouped scores, the computation of the median is essentially a process of counting scores arranged in order until the middle score is reached.

ARITHMETIC MEAN

Another frequently used representative measure of a group of scores is the arithmetic mean, often referred to as the mean or the average. In everyday affairs, we hear and read such phrases as average income, mean annual rainfall, average number of runs batted in, average yards gained from the line of scrimmage, and average class size. In each of these instances, the arithmetic mean is the single measure used to represent a group of measures.

Everyone is probably familiar with the computation of an average or mean. The mean of a group of scores is defined as the sum of the scores divided by the number of scores. For example, to find the arithmetic mean of the following scores: 12, 23, 15, 30, 8, 6, 18:

	12			
First, find the	23	Then, divide		The result
total sum of the	15	the total by		is the
scores.	30	the number of	$\frac{112}{7} = 16$	arithmetic
	8	scores in the		mean.
	6	group.		
	18			
Total	<u>112</u>			

This definition is particularly applicable to a set of raw, ungrouped scores. Since we are interested only in the total sum of the scores, it is not necessary to arrange the scores in order of size as was necessary for the median.

In setting up the definition of the arithmetic mean symbolically, if we let

X represent each of the raw, ungrouped scores

Σ , the Greek letter sigma, represent "the sum of" or "summation of"

N represent the total number of scores in the group

and $\bar{X}$ (X -bar) represent the arithmetic mean

then, $\bar{X} = \frac{\Sigma X}{N}$

In the example, X represents each of the raw scores 12, 23, 15, 30, 8, 6, 18; N is 7; ΣX equals 112; and $\bar{X}$ equals 16, the arithmetic mean of the seven scores.

Next, it is necessary to define what is meant by the "deviation" of a score from the mean. If we subtract the mean from any score, the result is the *deviation* of that score from the mean. Symbolically, if we let

X represent the raw score

$\bar{X}$ represent the arithmetic mean

and d represent the deviation of the score
from the mean

then, $d = X - \bar{X}$

If the mean of a distribution is equal to 22 and we wish to find the deviation of a particular score in the distribution, say 29, then according to the definition, deviation $d = 29 - 22 = +7$ units. Similarly, the deviation of the score 17 is $d = 17 - 22 = -5$ units. The plus and minus signs are necessary to indicate whether the deviations are above or below the mean.

Now let us refer to the 7 scores for which we computed a mean. This time we shall compute the deviation of each score from the mean ($\bar{X} = 16$).

Raw Score X	Deviation $d = X - \bar{X}$	Deviations Above Mean	Deviations Below Mean
12	- 4		
23	+ 7	+ 7	- 4
15	- 1	+14	- 1
$\bar{X} = 16$ 30	+14	+ 2	- 8
8	- 8	—	-10
6	-10	+23	—
18	+ 2		-23

If we add the deviations of the scores above the mean, we obtain +23. Similarly, if we add the deviations of the scores below the mean, we get -23. And if we add the two sums, $+23 + -23 = 0$, we get zero. This is a very important property of the arithmetic mean; namely, that the sum of the deviations of scores above the mean is equal to the sum of the deviations of scores below the mean. In other words, the sum of the deviations of all the scores from the mean is zero. If you wish actually to observe this property, take a ruled stick and place small blocks of equal weight at positions on the stick that correspond to the seven scores, 12, 23, 15, 30, 8, 6, and 18, as in Figure 29.

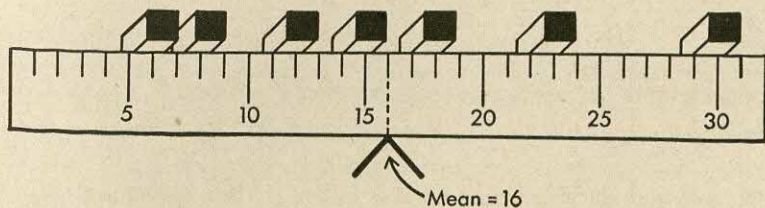


Figure 29. Mean as a point of balance.

The stick will balance as indicated if supported at the mean. Thus the arithmetic mean is representative of a group of scores in the same manner as the center of gravity is representative of all the parts of a rigid body.

The fact that the sum of the deviations of scores from the mean is zero provides another method for computing the mean. We can guess what the mean of a distribution is likely to be and then correct our guess for the true mean. This method can be shown by using again our seven scores.

	Raw Scores	Deviation (d') from
	X	Guessed Score 19
1. Start by guessing a mean. Here we guess that 19 is the mean.	12	- 7
	23	+ 4
	15	- 4
2. Find the deviation (d') of each score from 19.	30	+11
	8	-11
	6	-13
	18	- 1
3. Add the deviations and divide by the number of deviations. This is the <i>correction</i> .		Correction =
$\frac{(\Sigma d')}{(N)}$		$\frac{\Sigma d'}{N} = \frac{-21}{7} = -3$
		$\Sigma d' = -21$
		True Mean = Guessed Score + Correction
		$= 19 + (-3) = 16$
4. Add the correction to the guessed score to obtain true mean.		

For ungrouped scores, this method has no particular advantage over that of simply adding the scores and dividing by the number of scores. But it is the conventional method for scores grouped in intervals and here it does save time and energy.

To compute a mean from a large number of ungrouped scores, many of which are duplicated, we can use a computing machine and apply the first formula, $\bar{X} = \frac{\Sigma X}{N}$. Or, if none is available, we may use a slightly different formula and proceed as follows:

Score	Frequency	
X	f	fX
13	3	39
12	9	108
11	17	187
10	32	320
9	24	216
8	19	152
7	11	77
6	5	30
	$N = 120$	$\Sigma fX = 1129$

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{1129}{120} = 9.4$$

As indicated above, multiply each score by its frequency and then find the total of these products (ΣfX). Divide this sum by the total number of scores ($\frac{\Sigma fX}{N}$) and the result is the mean of the scores. For scores in a frequency

table, the original definition ($\bar{X} = \frac{\Sigma X}{N}$) now becomes $\bar{X} = \frac{\Sigma fX}{N}$, where f is the frequency of each of the scores.

We are now ready to learn the computation of the mean for scores grouped in intervals. In obtaining a mean from grouped scores, we combine the process of correcting a guess and of multiplying a score by its frequency. We need to assume that the mid-point of each interval represents the scores in the interval.

Using the same scores for which we obtained a median (Table 5), we say that our guessed mean (X_o) is the mid-point of the interval 60–62 or 61. Our table and calculations should look like this.

Mid-point Deviation Frequency					
Interval	X	$d' = (X - X_o)$	f	fd'	
72-74	73	+12	2	+24	
69-71	70	+ 9	13	+117	
66-68	67	+ 6	21	+126	
63-65	64	+ 3	44	+132	+ 399
60-62	61	0	28	0	
57-59	58	- 3	19	- 57	
54-56	55	- 6	7	- 42	
51-53	52	- 9	3	- 27	
48-50	49	-12	1	- 12	- 138
				$N = 138$	$+261 = \Sigma fd'$

True Mean = Guessed Mean + Correction Correction =

$$\bar{X} = X_o + \frac{\Sigma fd'}{N} = 61 + 1.9 = 62.9$$

$$\frac{\Sigma fd'}{N} = \frac{+261}{138} = + 1.9$$

In the deviation column (d') of the above table, we note that each deviation may be divided by 3, which is the size of each of the intervals. If this is done, then the (d') column, which formerly looked like this: +12, +9, +6, +3, 0, -3, -6, -9, -12, now becomes this: +4, +3, +2, +1, 0, -1, -2, -3, -4. This way of expressing deviations, in effect by counting intervals, is conventional and simplifies our calculations. We shall symbolize such deviations by the small letter u . When we use u , we must later multiply Σfu by the size of the interval so as to have the proper correction.

The usual computation of the mean is made from grouped scores, uses the guessed mean and correction method, and counts interval deviations rather than score deviations. This computation, the one you should use, is described

in Table 6. Note that the term "assumed mean" is used here rather than "guessed mean." The two mean the same thing and may be used interchangeably.

TABLE 6
An Outline for Computing the Mean from
the Formula $\bar{X} = X_o + \frac{i\Sigma fu}{N}$

Steps to follow:	Interval	<i>f</i>	<i>u</i>	<i>fu</i>	
1. Select an interval whose mid-point will be the assumed mean. Here we select the interval 60-62, and its mid-point, 61, is the assumed mean (X_o).	72-74	2	+4	+ 8	
	69-71	13	+3	+39	
	66-68	21	+2	+42	
	63-65	44	+1	+44	+ 133
	60-62	28	0	0	
2. Set up the <i>u</i> -column in terms of the number of intervals above and below the selected interval 60-62.	57-59	19	-1	-19	
	54-56	7	-2	-14	
	51-53	3	-3	- 9	
	48-50	1	-4	- 4	- 46
		<u><i>N</i> = 138</u>		<u>$\Sigma fu = + 87$</u>	
3. Multiply each <i>f</i> by <i>u</i> and then find the sum of the column (Σfu).					
4. Multiply this sum (Σfu) by the size of the intervals ($i = 3$) and then divide by <i>N</i> . This is the correction $\left(\frac{i\Sigma fu}{N}\right)$.					
Correction = $\frac{i \cdot \Sigma fu}{N} = \frac{3 \cdot 87}{138} = \frac{261}{138} = +1.9$					
5. Add this correction to the assumed mean and the result is the true mean. $\bar{X} = X_o + \frac{i\Sigma fu}{N}$					
$\bar{X} = X_o + \frac{i\Sigma fu}{N} = 61 + 1.9 = 62.9$					

The histogram may be used to illustrate the position of the mean as it was to illustrate the median. You will recall that the median is the point on the histogram that divides the area in half. Now, if we visualize the histogram as being cut out of cardboard or some other stiff material having weight, then the mean is the point on the base on which the histogram balances itself. This is illustrated in Figure 30, in which the histogram is drawn from the frequency distribution in Table 6.

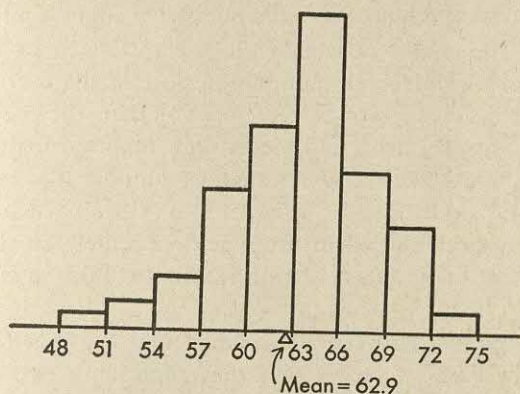


Figure 30. Illustrating the mean as the center of balance of the histogram.

COMPARISON OF THE MEDIAN AND MEAN

Although the median and the mean are both representative scores of a group of scores, we can see now that the definitions and the processes for finding them are quite different. The median is sought by a process of counting ranked scores until the middle score is reached. The mean is found through a process of determining the balance point of the scores. These differences have implications in the application of the mean and median, which we shall now discuss.

In the first place, if the distribution of scores is symmetrical or nearly symmetrical then there is little or no difference between the middle point and the balance point of the distribution. Consequently, the mean and median would have about the same numerical value, and in this sense there would be no choice between the two measures.

For skewed distributions, on the other hand, the numerical values of the mean and median will ordinarily be quite different, and we shall discuss such cases with the aid of Figure 31.

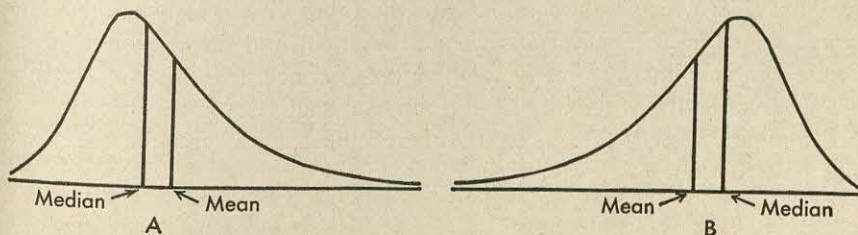


Figure 31. The effect of skewed distributions on the mean and median.

As illustrated, the mean is "pulled" toward the end of the distribution in which there are a few extreme scores, while the median continues to divide the area in half. The choice of whether to use the mean or the median for skewed distributions will depend upon the purpose of the user. If it is desired to avoid any undue influence of extreme scores, then the median should be preferred. Often both the mean and median are reported to indicate whether or not the distribution is skewed. For example, suppose it is reported that the mean is equal to 47 and the median is equal to 55 for a particular distribution. Then we would suspect that there are a few extremely small scores in the distribution and that the graph of the distribution would be similar to Figure 31-B.

A further distinction between the median and the mean must be made with respect to the types of scores they may represent. You may recall that in the first chapter we presented some measurement symbols that expressed scale position and others that expressed rank position. In the calculation of the mean it is necessary either to add the scores or to determine deviations of the scores, and both of these processes require unit-scale numbers. As a result, the mean should be used *only* for distributions of unit-scale scores and would be meaningless if used with scores expressing rank or order. The median, on the other hand, may be used with either type of measurement symbol and therefore it is the most widely applicable descriptive measure of central tendency. From the standpoint of theoretical statistics, the mean has certain advantages since it figures in the computation of other statistical measures such as the standard deviation, the correlation coefficient, and the analysis of variance.

Describing the Variability of a Group

It should take only a moment's reflection to realize that to describe a group of measures by a single representative measure would be incomplete. There certainly would need to be some indication of the spread, or dispersion, or variability of the scores in the group. To know, for example, that the average annual temperature in two American cities is the same and equal to 70° would still leave one to wonder about the climate of the two cities. In one city the temperature could range from extremely cold to extremely hot, and still have an average temperature of 70°. In the other city, which has the same average temperature, the variation could be small and hence would have a more favorable climate as far as temperature is concerned. Of many possible ways to describe the variability of a group, we shall discuss the three indexes most commonly used, the range, interpercentile ranges, and the mean and standard deviation.

RANGE

Of these, the simplest and crudest method is the *range*. The range of a distribution of scores is defined as the difference between the highest score and the lowest score in the distribution. To see how the range is computed,

look at the raw scores presented in Table 3. There we find that the highest score is 133 and the lowest 91. Therefore, by definition the range is equal to 42 ($133 - 91$). The range, then, presents us with an approximate basis for comparing the variability of two distributions as in the case of the two cities discussed above. If the range of temperatures in one city is 93 and the range in the other city is 47, this indicates that the variation in the latter city is less. The range is only a rough indication of variability because it is conceivable that two distributions may have the same or nearly the same range and yet have different variabilities, as in Figure 32.

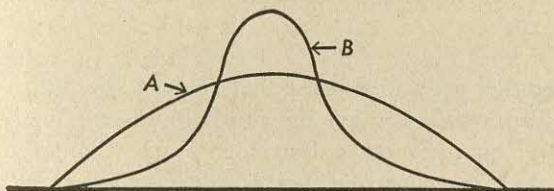


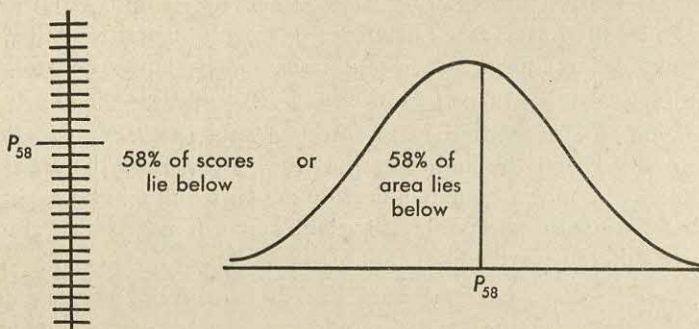
Figure 32. Two distributions having the same range but different variability.

In distribution *A*, the scores are scattered over the entire range while in distribution *B* most of the scores tend to cluster more closely around the center. Yet the highest and lowest scores are the same, making the ranges equal.

INTERPERCENTILE RANGES

From the previous discussion, we can see that the range is a function of the scores at either end of the distributions. In order to exclude these few extreme scores that determine the total range, a special type of range called the *interpercentile range* has been developed as a measure of variability. However, before we can define an interpercentile range, we need to know the meaning of a percentile.

Since percentiles will be discussed in complete detail in a later section dealing with measures of relative position, we shall develop here only what is necessary for an understanding of an interpercentile range. Briefly, a percentile is a score or point on a scale of scores below which a certain percentage of the total number of scores lie. If exactly 35 per cent of the total number of scores are less in value than the score of 52, then the score 52 is called the 35th percentile and it is designated as P_{35} . Likewise, if Johnny weighs 127 pounds and he is heavier than 80 per cent of his class at school, then his weight, 127 pounds, is the 80th percentile for that particular group, and is designated as P_{80} . In general, if N per cent of the total number of scores or N per cent of the area of a frequency curve lies below a certain value, then that value is called the N th percentile and is designated as P_n . The 58th percentile, for instance, could be schematically presented as follows:



Certain percentiles are called by special names. The 50th percentile, P_{50} , is called the median, which we have previously discussed. The 25th percentile is called the lower quartile and is designated as Q_1 . Similarly the 75th percentile is called the "upper quartile" and is designated as Q_3 . The 10th, 20th, 30th percentiles are called 1st, 2d, 3rd, . . . deciles respectively.

The *interpercentile range* may now be defined as the range or difference between any two symmetrically placed percentiles above and below the median. The range from the 10th to the 90th percentiles, $P_{90}-P_{10}$, or perhaps the range from the 7th to the 93rd percentiles, $P_{93}-P_{07}$, would serve to indicate the spread of the scores and yet cut off the few scores at either end of the distribution which may be erratic. The range from the 25th percentile to the 75th percentile, or in other words, from the lower quartile to the upper quartile, Q_3-Q_1 , is called the *interquartile range* and if the interquartile range is divided by 2, $\frac{Q_3-Q_1}{2}$, it is called the *quartile deviation*. In any case, an interpercentile range, interquartile range, or quartile deviation as a measure of variability is used in connection with the median as a representative score.

MEASURES OF VARIABILITY BASED ON DEVIATIONS

The third method of indicating the dispersion of scores is based upon the deviations that the scores have from the mean. Since these deviations are indications of distances of the scores from the mean, an average of these deviations would seem a natural way of indicating the spread of these scores.

We have already observed that the algebraic sum (taking into account the plus and minus signs) of the deviations from the mean is always zero. So the matter of finding the average of the deviations is not without difficulties. Some method must be found of preventing the positive and negative deviations from canceling out and still keeping the relative sizes of the deviations intact. Two methods are possible: to disregard the signs, or to square the deviations. The first method results in the *mean deviation* and the second method in the *standard deviation*.

Mean Deviation. First, consider the method of disregarding the signs.

Whenever the sign of a number is disregarded, the result is the "absolute value" of the number. Thus the absolute value of -3 is 3 and likewise the absolute value of $+3$ is 3 . The absolute value is symbolized by two vertical bars on either side of the number. $|N|$ means the absolute value of N . Now the *mean deviation* is equal to the sum of the absolute values of the deviations of the scores divided by the number of scores. This measure is not seen very often, but does provide a direct approach to indicating variability based on the deviations of scores.

Standard Deviation. The second method, squaring the deviations in order to get rid of the signs, provides the basis for computing the standard deviation. If the deviations of the scores are squared, the mean of these deviations is called the *variance* (s^2), and the square root of the variance is known as the *standard deviation* (s). The following example outlines the computation of the standard deviation for 7 scores whose mean is 16.

Steps to follow:

Steps to follow:

	<i>X</i>	<i>d</i> = <i>X</i> − $\bar{X}$	<i>d</i> ²	
1. Determine the deviation (<i>d</i>) of each score from the mean.	23	+7	49	$\bar{X} = 16$ $N = 7$
	12	−4	16	
	6	−10	100	
2. Square each deviation (<i>d</i> ²) and add the column, Σd^2	30	+14	196	
	18	+2	4	
	8	−8	64	
3. Divide the sum of the deviation squared column by the number of scores, $\frac{\Sigma d^2}{N}$, and the result is the variance, <i>s</i> ² .	15	−1	1	
		$\Sigma d = 0$	$430 = \Sigma d^2$	
		$\text{Variance } s^2 = \frac{\Sigma d^2}{N} = \frac{430}{7} = 61.4$		
4. Take the square root of the variance and the result is the standard deviation, <i>s</i> .		$\text{Standard deviation } s = \sqrt{\frac{\Sigma d^2}{N}} = \sqrt{61.4} = 7.8$		

In engineering, the standard deviation is called the root mean square (rms) which is a more descriptive term since the standard deviation is the square root of the mean of the squared deviations. In the example above, if the raw scores had a frequency other than 1, the formula for the standard deviation would be $s = \sqrt{\frac{\Sigma fd^2}{N}}$.

With this basic definition of the standard deviation, we can now turn to its computation from a frequency table. Here the procedure is essentially the same as the for the mean. First, we compute the standard deviation of the scores from the assumed mean, $\frac{\Sigma fd^2}{N}$, and then subtract the correction for the

true mean $\left(\frac{\sum fd}{N}\right)^2$. When this is carried out, the result then becomes

$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$. In case the deviations are in terms of interval units

(u), then the formula becomes $s = i \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}$ where i is the size of the interval. Table 7 outlines the computation of the standard deviation for scores grouped in intervals.

TABLE 7

An Outline for the Computation of the Standard Deviation

Steps to follow:	Interval	f	u	fu	fu^2
1. Assume a mean which is the midpoint of an interval and set up the u column.	80-84	3	3	9	27
	75-79	7	2	14	28
	70-74	5	1	5	5
	65-69	11	0		0
2. Multiply each f by its u and add the column, $\sum fu$.	60-64	16	-1	-16	16
	55-59	10	-2	-20	40
	50-54	8	-3	-24	72
3. Multiply each fu by its u and add the column, $\sum fu^2$.		$N = 60$	$\sum fu = -32$		$188 = \sum fu^2$
4. Substitute the values in the formula for standard deviation, $\sum fu = -32$, $\sum fu^2 = 188$, $i = 5$, $N = 60$.	$s = i \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 5 \sqrt{\frac{188}{60} - \left(\frac{-32}{60}\right)^2}$ $= 8.5$				

COMPARISON OF STANDARD DEVIATION AND INTERPERCENTILE RANGE

Although the interpercentile range and the standard deviation both give an indication of variability, we can see from the previous discussion that they are different in nature. The interpercentile range represents an actual range or distance from one point in the distribution to another point, and the percentage of scores between these two points can be stated. For example, the interquartile range, which is a special case of the interpercentile range, establishes two scores, the upper quartile and lower quartile, that include the middle 50 per cent of the distribution. The standard deviation, on the other hand, is simply a representative deviation of all the deviations of the scores, and, except in such theoretical distributions as the normal distribution, the standard deviation does not represent a range that includes a certain percentage of the scores. Thus the interpercentile range is a better descriptive measure than the standard deviation because it is more easily visualized and comprehended.

It has already been noted that the interpercentile range should be used

in connection with the median and that the standard deviation should be used with the mean. Consequently, the interpercentile range is particularly adaptable to measurement symbols that indicate rank position, and the standard deviation is limited to measurement symbols that indicate scale position. Like the mean, the standard deviation is useful in the computation of further statistical measures and in the area of sampling statistics. Therefore, from the standpoint of statistical analysis the standard deviation is preferred over the interpercentile range.

The interpercentile range and the standard deviation are equally difficult to interpret. For example, suppose we calculated that the interquartile range of a distribution is 17 and that the standard deviation is 9. What do these measures mean? Standing alone, these measures mean very little. They can be interpreted only by comparison with other results in similar situations. If it is found, say, that the interquartile ranges for other similar distributions is less than 17, then we would suspect that our distribution is somewhat dispersed. In general, any judgment regarding the degree or extent of variability of a distribution is a relative matter.

Describing the Relative Position of Individuals in Groups

We turn now from describing groups to describing the individuals within the groups and in particular the relative position of the individuals. As was indicated in previous sections, any score or measure standing by itself has little meaning. Just what does Mary's test score of 55 signify if that is all we know? In this section we shall discuss certain statistical procedures able to convert individual raw test scores to measures that indicate more meaningful relative position.

RANK ORDER

As usual, we shall start with the simplest technique for indicating relative position, that of establishing rank order. This method consists simply of arranging the raw scores in order of size and then of assigning a rank to each, the top score receiving Rank 1, and so on down the line. If a score is obtained by two or more persons, then the rank of the score is repeated the necessary times and the next score in line would have a rank which takes into account the previous repetition. For example, if three persons received a score that ranks third, then the ranks would proceed as follows: 1, 2, 3, 3, 3, 6, 7, etc. Note that the third rank is repeated the necessary three times, followed by rank 6, which indicates that the next score is 6th in line. Obviously, if the number of scores is large, the process of ranking will become unwieldy. Moreover, to know just the rank of a score is not sufficient because the same rank would have different meaning in groups of different sizes. For example, consider a rank of 15 in a group of 432 scores and then consider a rank of 15 in a group of 20 scores.

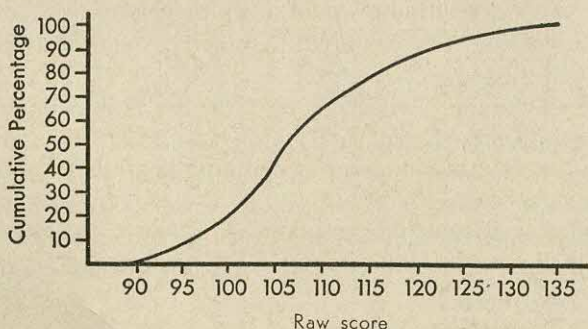
PERCENTILE RANK

To correct the limitations of simple rank order, another measure of relative position has been developed, the percentile rank. Percentiles have been discussed previously in Chapter 1 in connection with measurement symbols that indicate order position, and earlier in this chapter in connection with interpercentile ranges. Here we shall discuss the percentile in connection with

TABLE 8

An Outline for Developing a Cumulative Percentage Curve

Steps to follow:		<i>f</i>	<i>Cumulative frequency</i>	<i>Cumulative percentage</i>	
1. Start at the bottom of the frequency column and begin adding the frequencies one at a time, recording the result of each addition in the cumulative frequency column.	130-134	1	90	100	
	125-129	3	89	99	
	120-124	6	86	96	
	115-119	10	80	89	
	110-114	14	70	78	This is the
	105-109	21	56	62	percentile
	100-104	19	35	39	rank of
	95- 99	11	16	18	109.5
2. Divide each number in the cumulative frequency column by $N = 90$ and record the result in the cumulative percentage column. Thus each number in the cumulative percentage column represents the percentile rank of a boundary point.	90- 94	5	5	6	
	$N = 90$				
3. Set up a table of boundary points and percentile ranks and plot from this table the cumulative percentage curve.	<i>Boundary point</i>	<i>Percentile rank</i>			
	89.5	0			
	94.5	6			
	99.5	18			
	104.5	39			
	109.5	62			
	114.5	78			
	119.5	89			
	124.5	96			
	129.5	99			
	134.5	100			



determining the relative standing of an individual raw score in a group of scores.

The procedure for finding the percentile rank of a raw score in a small group of scores is simply that of finding the percentage of the total number of scores that fall below the given raw score. Suppose, for example, out of a total of 34 scores, 14 scores fall below the raw score 67, or approximately 41 per cent of the scores fall below 67. Then the percentile rank of the raw score 67 for this particular group is 41, which is designated as P_{41} .

Finding the percentile ranks of raw scores in large distributions grouped in intervals is a more involved process. In this case we can employ a bit of strategy to save ourselves prolonged computations. If you will look closely at a frequency table, you may notice that the computation of the percentile rank of a boundary point between two intervals is a very simple and direct procedure. For any boundary point, the percentile rank is found by adding the frequencies of the intervals below the boundary point and dividing the sum by the total frequency. Now, after finding the percentile rank of each boundary point, we have enough data to plot a graph from which the percentile rank of any score in the distribution may be determined. Of course this graph has a special name. It is called a *cumulative percentage curve* or an *ogive*. Table 8 presents an outline of the steps leading to the plotting of a cumulative percentage curve from which percentile ranks of individual raw scores may be found.

From the ogive not only can the percentile rank of each raw score be found but also the median, the quartiles, and the deciles, and these figures provide the basis for computing the interquartile range or any interpercentile range. If the curve is carefully plotted on graph paper, the readings can be sufficiently accurate for most purposes.

STANDARD SCORE

The third and final measure of a relative position to be discussed here is the standard score. The standard score of a certain raw score is equal to the deviation of the raw score from the mean divided by the standard deviation. Ordinarily the letter z is used to refer to the standard score and thus we often see the standard score referred to as the z score. The formula for the standard score is as follows:

$$\text{standard score } (z) = \frac{X - \bar{X}}{s}$$

Where $X - \bar{X}$ is the deviation of the score from the mean and s is the standard deviation.

Suppose that we wish to find the standard score of the raw scores 52 and 40 where the mean and standard deviation of the distribution containing them are: $\bar{X} = 43$

$$s = 6$$

then $z = \frac{52 - 43}{6} = \frac{+9}{6} = +1.5$ for the raw score of 52

and $z = \frac{40 - 43}{6} = -.5$ for the raw score of 40

Thus the standard score expresses the relative position of the raw score by indicating the number of units of standard deviation between the score and the mean. A standard score of $z = +1.5$ would indicate that the score is one and a half standard deviations above the mean. A standard score of $z = -.5$ would show that the score is half a standard deviation below the mean. Standard scores larger than $z = +3$ or less than $z = -3$ are not likely to occur unless the distribution is badly skewed.

COMPARISON OF PERCENTILE RANKS AND STANDARD SCORES

Both the percentile ranks and standard scores are widely used as measures of relative position, with each having its own particular advantages and disadvantages. The percentile rank is readily understood by most of us at first encounter, but percentiles are not evenly spaced throughout the distribution and they are of no use for further statistical computations. Standard scores are evenly spaced throughout the distribution and they are used in further statistical work. They have the disadvantage of not being easily understood. Moreover, their use is limited to unit-scale raw data and it is difficult to interpret them for markedly skewed distributions. So, once again, the selection of the type of measure to be used depends upon the purpose of the user and the nature of data at hand.

NORMS

Up to this point, nothing specific has been said about the groups of students upon which all these statistical manipulations have been made. It has been heretofore tacitly assumed that the raw scores came from any arbitrary group that the teacher may encounter. This is not always the case, however. Attempts are often made to obtain raw scores from carefully selected groups that are to represent a larger population. Using various sampling techniques, for example, a group of 200 students may be selected, which is representative of all the fifth-grade students in a particular region. The arithmetic means, medians, percentile ranks, and standard scores computed from raw data obtained from such representative groups are called *norms*.

The interpretative value of having norms available is evident and has already been indicated in chapters 1, 3, and 6. Without norms, a teacher can compare the performance of a student only with the other members of his local class. With norms, however, a teacher has a much broader basis for comparison.

There are two basic methods of relating an individual score to a norm.

1. A sequence of norms may be established representing a gradation of

level, and the individual score is matched with the closest norm to determine the level of performance or status indicated by the score.

2. The relative standing of the individual score is determined directly with respect to the representative group.

The first method has led to the development of age and grade norms, and the second method is the basis for establishing percentile rank and standard score norms.

Age and Grade Norms. These norms are used whenever the variation of a dimension is closely related to age or grade in school. Such dimensions would be achievement in various subjects, height, intelligence, reading comprehension, and vocabulary.

Age norms are often used for height. The mean heights of five-year-old boys, six-year-old boys, and so on, are computed, thus establishing a sequence of height norms for boys by age groups. The height of a boy then can be compared with these norms. For instance, it may happen that a five-year-old boy has a height typical of six-year-old boys. Mental age is another example of age norms. An intelligence test is administered to various age groups and the mean or median is computed for each age level. These mean or median scores on the intelligence test become age norms or mental age scores. Consequently, the score of an individual pupil may be used as a basis for determining his mental age.

Scores on various achievement tests and reading tests are often converted to norms based upon grade level in school, in the same fashion as for age norms. Means or medians are computed for each grade-level group of students and thus are called grade norms. The score of an individual pupil on one of these tests is then compared with the grade norms to determine the grade level of his performance.

Percentile Rank and Standard Score Norms. These norms are obtained directly from a single representative group. A test is administered to this representative group, and the percentile ranks or standard scores computed for this group become the percentile norms or standard score norms as the case may be. For this method, the individual score of a student on a test is converted directly to relative standing in a group typical of the total population of which the student is a member. As an example, the score of a fifth-grade pupil on a reading test is converted to a percentile rank or standard score in terms of the total population of fifth-grade pupils. This procedure is different from the first procedure, where the pupil's score was used as a basis for determining the group of which the pupil would be typical.

Tests developed on the basis of data obtained from groups of students that are typical of much larger groups of students are commonly called *standardized tests*. (See Chapter 6, pages 121–126) Hence, by the very nature of their construction, all standardized tests should provide norms, either of the age and grade type or of the percentile and standard score type. Ordinarily these norms are found in tables contained in the test manuals where there is

provided a column of raw scores and alongside is a column of their norm equivalents. If a test is appropriate to more than one group, we should expect to find a separate set of norms for each group.

Precautions in the Use of Norms. In the interpretation and use of norms, certain precautions should be kept constantly in mind.

1. Norms do not represent what is ideal. Norms are useful only in describing the level of performance of a student. In other words, a norm is a measurement symbol and not an evaluative symbol (see Chapter 9, pages 190-195).

2. Norms do not indicate that a pupil should be advanced or retarded to the age or grade group in which the pupil's score is typical. If the height of a seven-year-old child is typical of the height of six-year-olds, this does not mean that the child should be six years old rather than seven years old. If a fourth-grade pupil's achievement level in a certain subject is typical of sixth-grade pupils, this does not necessarily mean that this pupil should be advanced to the sixth grade. The high level of performance of this pupil may be due primarily to superior mastery of fourth-grade material and not to ability to do sixth-grade work.

3. The problem of obtaining a truly representative sample of a given population is so great that it seriously hampers the valid use of norms. Test experts are now more aware than ever of the inadequacy of present methods in obtaining representative samples for establishing test norms. Consequently, teachers should be extremely cautious in applying norms to their students. Teachers should at least closely examine the description of the group on which the test was standardized to see if it is comparable to the group they are working with.

4. The teacher should make certain, before applying norms, that the test is not in any way unfair to the particular group at hand. It has happened that a certain test tends to discriminate against certain culture groups or contains items on topics which were omitted by the teacher in class.

5. Although norms are based upon age, grade, and standard scores which are scale position symbols, norms do *not* indicate a unit scale of performance and therefore should not be treated as scale position symbols. At best, norms are still measures of relative position.

Summary

In this chapter we have presented some basic descriptive statistical operations. The significance of the operations is to organize raw measurement data into more meaningful data, thus to accomplish or further facilitate precise appraisal of the phenomenon in question. These statistical concepts and procedures have these essential purposes:

1. To provide an over-all picture of the group pattern
2. To provide a representative score for a group of scores

3. To provide a measure of the variability of a group of scores
4. To provide a measure of relative position for individual raw scores

In connection with the first purpose, visualizing the group pattern of scores, the following statistical devices were discussed: (a) The arrangement of the scores in order of size, (b) the establishment of score intervals leading to the setting up of a frequency table, (c) the construction of a bar graph called the histogram, and (d) the plotting of the frequency polygon, which can be used to approximate a smooth frequency curve. The second purpose, providing a representative score, led to a discussion of the mode, median, and mean. The measures of variability presented were the range, the interpercentile range, the interquartile range (which is a special case of the interpercentile range), the mean deviation, the variance, and the standard deviation. Finally, as measures of the relative position of an individual in a group, we examined rank order, percentile rank, and the standard score. Norms based upon representative groups are a further aid in interpreting the relative positions of individuals.

For each of the four purposes, the available statistical measures range in refinement from crude approximations to precise and refined expressions. As these statistical devices are encountered and used, it is important that we have in mind their limitations, their special purposes, and the assumptions that underlie them.

EXERCISES

1.	68	75	78	77	82	80	71	90	86	68
	76	78	73	88	85	72	78	67	80	70
	61	77	71	76	93	76	76	66	73	76
	86	60	89	73	84	81	84	87	66	81
	81	84	73	64	79	77	88	82	52	72
	75	77	74	82	79	71	75	82	77	80

Using the above raw scores:

- a. Construct a frequency table.
 - b. Construct a histogram and frequency polygon.
 - c. Compute the mean.
 - d. Compute the median.
 - e. Compute the standard deviation.
 - f. Compute the interquartile range.
 - g. Draw an ogive.
2. What criteria would you use for judging the suitability of a frequency table?
 3. What statistical techniques are appropriate for use with raw scores obtained from teacher-made tests?
 4. Give examples of the misuse of statistics in educational measurement.

5. Examine the manual of a standardized test and summarize the procedures used in establishing norms for the test. Make a critique of these methods.

BIBLIOGRAPHY

1. Adams, J. K., *Basic Statistical Concepts*. New York: McGraw-Hill Book Co., 1955.
2. Clark, Charles E., *An Introduction to Statistics*. New York: John Wiley & Sons, Inc., 1953.
3. Croxton, F. E., and Cowden, D. J., *Applied General Statistics*. New York: Prentice-Hall, Inc., 1940.
4. Freund, John E., *Modern Elementary Statistics*. New York: Prentice-Hall, Inc., 1952.
5. Guilford, J. P., *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill Book Co., Inc., 1950.
6. Johnson, P. O., *Statistical Methods in Research*. New York: Prentice-Hall, Inc., 1949.
7. Sprowls, R. C., *Elementary Statistics for Students of Social Science and Business*. New York: McGraw-Hill Book Co., 1955.
8. Walker, Helen W., *Elementary Statistical Methods*. New York: Henry Holt & Co., Inc., 1943.
9. Wilks, S. S., *Elementary Statistical Analysis*. Princeton: Princeton University Press, 1951.

CHAPTER 8

FURTHER STATISTICAL CONCEPTS IN MEASUREMENT

In the preceding chapter, we discussed certain basic statistical concepts that describe in various ways the measurement data we may have, and we developed outlines for computing these descriptive devices. In this chapter we shall consider a few additional statistical concepts often encountered in educational measurement, which stem from two important statistical terms, the normal probability curve and the correlation coefficient. The normal probability curve is often called the "normal curve" and is basic to any consideration of such topics as sampling, confidence limits, and standard error, discussed later in this chapter. The correlation coefficient describes the co-relationship between paired measures—the degree to which one variable is related to another variable. The correlation coefficient is basic to an understanding of the technical meaning of reliability and validity as used in measurement, which also will be discussed later. The computation of these statistical measures will not be outlined. Only the definitions of the terms will be developed and some discussion of their application will be made.

Normal Probability Curve

As its name implies, the normal probability curve is based upon the idea of probability. Therefore, any definition of this curve must first start with some basic notions of probability. After having lived awhile, our intuitive conception of probability has been pretty well formed. Our experience has shown us that certain situations involve more risk than other situations. The probability of our being involved in an auto accident on a road with heavy traffic is much greater if we were driving at a speed of 80 miles per hour than if we were driving at 20 miles per hour. We also know that the probability of our living for the next ten years generally decreases with our age.

The above examples, however, are rather complex, so let us confine ourselves to situations where the probabilities are easily computed. For instance, suppose we have a container that has in it 30 white balls and 20 black balls, what is the probability of drawing a black ball? The answer would be obviously, $20/50$ or 40 per cent. If we threw a single die, what is the probability

of throwing a 3? Since only one of 6 possible ways for a single die to turn up would be a 3, the probabilities of throwing a 3 is $\frac{1}{6}$ or $16\frac{2}{3}$ per cent.

THE PROBABILITY RATIO

Probability is generally expressed as a ratio, and we shall formalize the process of determining probability as follows. If of n equally likely possibilities m of these are favorable to the happening of a certain event, then the probability that the event will happen is the ratio m/n .

Another simple example of probability is that of coin-tossing. Suppose we have three coins, what is the probability of tossing two heads? To answer this, we must first consider all the "equally likely" possibilities in which three coins may fall. They are enumerated as follows: H (heads), T (tails)

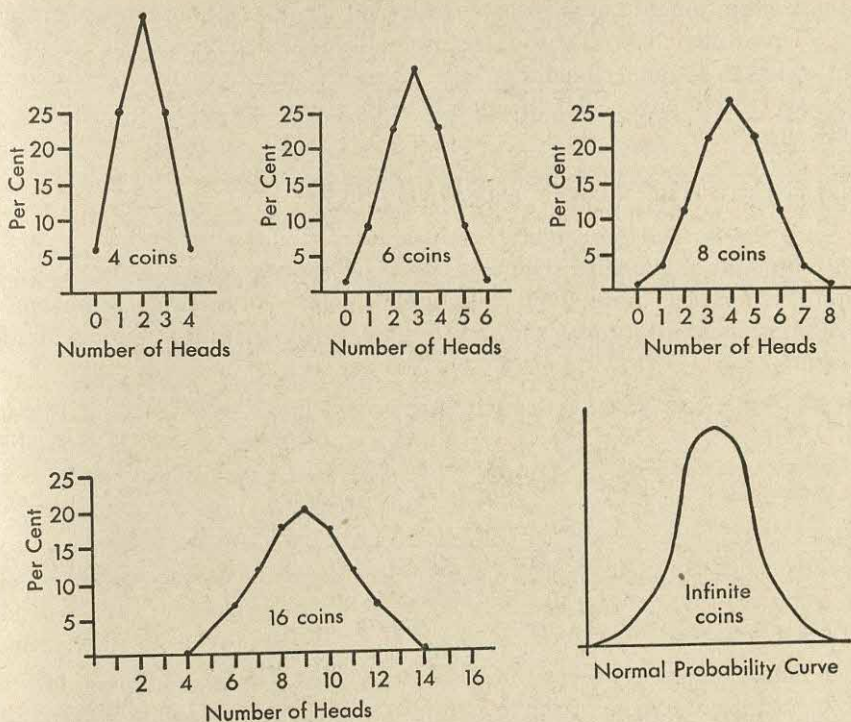
- | | | | |
|-----|-------|-----|-------|
| 1. | H H H | *5. | T H H |
| *2. | H H T | 6. | T H T |
| *3. | H T H | 7. | T T H |
| 4. | H T T | 8. | T T T |

There is a total of eight "equally likely" ways in which three coins may fall and of these there are three ways (marked by an asterisk) in which the result will be two heads showing. Then by definition, in the tossing of three coins the probability of two heads showing is $\frac{3}{8}$. Likewise, the probability of getting three heads is $\frac{1}{8}$, of getting one head is $\frac{3}{8}$, and of getting no heads is $\frac{1}{8}$. Note that these are expected probabilities. The expected probabilities of number of heads appearing for different number of coins are shown in Figure 33.

Number of coins tossed	Probability of a given number of heads appearing							
	0	1	2	3	4	5	6	7
4	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$			
6	$\frac{1}{64}$	$\frac{6}{64}$	$\frac{15}{64}$	$\frac{20}{64}$	$\frac{15}{64}$	$\frac{6}{64}$	$\frac{1}{64}$	
8	$\frac{1}{256}$	$\frac{8}{256}$	$\frac{28}{256}$	$\frac{56}{256}$	$\frac{70}{256}$	$\frac{56}{256}$	$\frac{28}{256}$	$\frac{8}{256}$
16	$\frac{1}{65,536}$	$\frac{16}{65,536}$	$\frac{120}{65,536}$	$\frac{560}{65,536}$	$\frac{1820}{65,536}$	$\frac{4368}{65,536}$	$\frac{8008}{65,536}$	$\frac{11,440}{65,536}$

Figure 33. Probabilities in coin tossing.

The figures shown are the frequency polygons for the various numbers of coins, showing the distribution of the expected probabilities for the different number of heads appearing.



Since the values of the expected probabilities may be computed from a formula based on the expansion of the binomial $(\frac{1}{2} + \frac{1}{2})^n$ where n is the number of coins tossed, these frequency polygons are called binomial distributions.

DEFINITION OF THE NORMAL PROBABILITY CURVE

In the binomial distributions, it is noticed that the shape of the distributions become smoother and more bell-shaped as the number of coins increases. This observation leads us to a descriptive definition of the normal probability curve. *The normal probability curve is the ultimate shape of the distribution of expected probabilities of number of heads showing when the number of coins tossed is increased without limit.*

The purpose of this brief development has been to illustrate how the normal probability curve is related to probability. One can readily see that it is a very special curve. It is *not* a curve obtained from measurement data like the curves obtained in the previous chapter. It is a theoretical distribution, represented by the mathematical equation $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Its usefulness lies in the special properties it has and in the fact that variations of certain phenom-

ena have been found to approximate closely the shape of a normal probability curve. Unfortunately, this curve has been ill-used in educational measurement owing to misunderstanding of the theoretical nature of the curve and it is for this reason that special attention is being given here to the normal probability curve.

PROPERTIES OF THE NORMAL PROBABILITY CURVE

Since the normal probability curve is a probability distribution, its area then represents probability. This is its special property, which makes it such a useful theoretical curve. First, take a brief look at the normal probability curve in its standard form as shown in Figure 34. We note that the z scale or

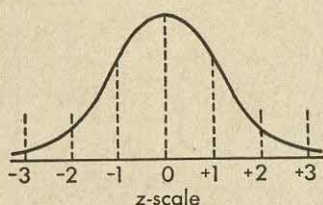


Figure 34. Normal curve and z -scale relationship.

standard scores, which were discussed in the previous chapter, are used. Zero is placed at the mean and the scale goes to the right and left of the mean in terms of standard deviation units. Thus, on this scale, -2 means two standard deviations below the mean, and $+1$ means one standard deviation above the mean. You will note that three standard deviations above and below the mean cover pretty much the scope of the curve.

Area— z Score Relationships. With the normal probability curve in standard form, the portions of areas between any pair of z scores is also standardized. For instance, it has been computed that 68 per cent of the total area under the curve lies between the z scores $+1$ and -1 , and further that 95 per cent of the total area lies between the z scores $+2$ and -2 . This is illustrated in Figure 35.

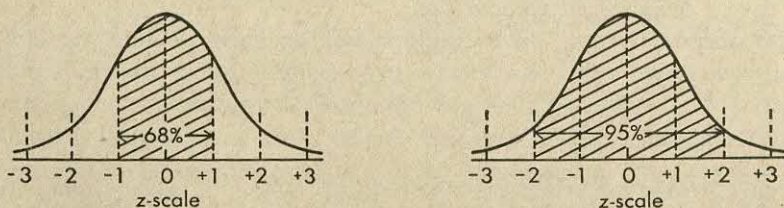
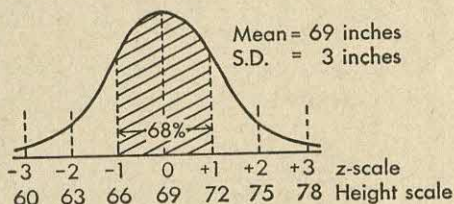


Figure 35. Portions of area of normal distribution contained between given z scores.

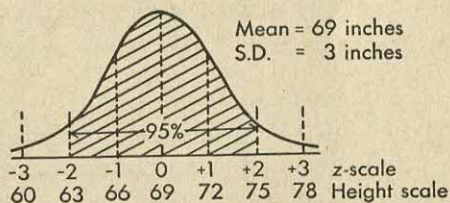
Area percentages between other pairs of z scores can be found by consulting a table for which these values have been computed.¹

z Score Probability Relationships. The relation between portions of areas under the curve and probability can perhaps best be shown by an example. Suppose we assume that the distribution of heights of men in the United States approximates the normal probability curve. Studies have shown that this is not an unreasonable assumption. We shall take as the mean 69 inches (5 ft. 9 in.) and as the standard deviation 3 inches, which are close to the results reported by these studies. Now we are in a position to make certain statements of probability.

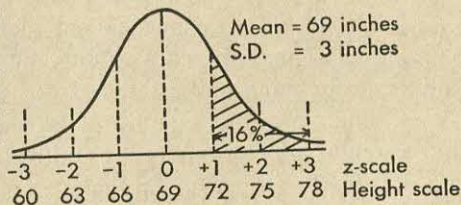
1. Subtracting three inches (one standard deviation) from, and adding three inches to, the mean of 69 inches gives us the interval 66 inches (5 ft. 6 in.) to 72 inches (6 ft.) and in this interval, 68 per cent of the distribution. Hence, if we choose a man at random, the probability is .68, or the chances are 68 out of 100, that his height is somewhere between 66 inches (5 ft. 6 in.) and 72 inches (6 ft.).



2. If we subtract and add two standard deviations (6 inches) to the mean, we get the interval 63 inches to 75 inches, which contains 95 per cent of the normal probability distribution. If we choose a man at random, the probability is .95, or 95 chances out of 100, that his height is somewhere between 63 inches (5 ft. 3 in.) and 75 inches (6 ft. 3 in.).



3. Since 16 per cent of the normal probability distribution lies above the z score +1, then we can also make the following statement: The probability is only .16, or 16 chances out of 100, that a man's height is over 72 inches (6 ft.). We might also say that the probability is .84, or 84 chances out of 100, that a man's height is less than 72 inches (6 ft.).



This same general approach may be applied to other situations involving the normal probability curve. For instance, studies have shown that the distribution of IQ's based on the Stanford-Binet scale approximates the normal

¹ See appendix, page 478.

probability curve. The mean IQ has been set at 100 and the standard deviation is equal to 16.² As shown in Figure 35, we can relate the normal probability curve to the distribution of IQ's and make certain probability statements.

1. The chances are 68 out of 100 that a person's IQ is somewhere between 84 and 116.
2. The chances are 16 out of 100 that a person's IQ is above 116.

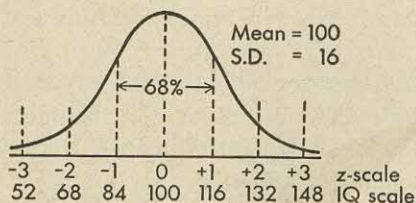


Figure 36. Probability of a person having a given IQ.

From these few simple examples we can see how probability can be obtained from the normal probability curve. Likewise, it becomes more obvious as to why it is desirable to work with distributions of data that approximate the curve of normal probability. As a matter of fact, the desirability of using the normal curve is so strong that it is often used when it should not be. This leads us to the question, for what types of measures can we reasonably assume the curve of normal probability to be applicable?

Applicability and Use of the Normal Probability Curve. Clues for answering the question concerning when the normal probability curve is applicable may be obtained by referring to the example of tossing coins. This example was used to explain the theoretical development of the normal curve earlier in this chapter. There were certain characteristics of this example that we might point out at this time.

1. There had to be a sufficient number of coins used before the distribution of number of heads approximated a normal distribution.

2. Each coin was independent of the other. In other words, whether a certain coin came up heads or tails did not depend in any way upon whether another coin came up heads or tails.

3. Each coin had equal importance in the final result. No one coin had more weight in determining the number of heads showing than any other coin.

4. Finally, for each coin there is a 50-50 chance that heads will show.

Such considerations are involved in why the distribution of the number of heads for tosses of coins resulted in a normal probability distribution. In general, the normal curve is applicable to distributions of measures that are the result of several definable factors, each of which is independent, has equal

² See Terman and Merrill, *Measuring Intelligence*, Boston: Houghton Mifflin Co., 1937, pp. 33-51.

weight, and for which there is a 50–50 chance of the factor's being present or not.

With this background, let us see which human traits are most likely to result in distributions approximating the curve of normal probability. Because of the "gene theory," we would expect a normal probability distribution for hereditary Mendelian traits in humans provided there is no selective factor or special environmental condition affecting the trait. The genes involved for each of these traits would be analogous to the coins in the previous example. Figuratively speaking, each person represents a "toss" of the genes with some genes dominant and others recessive, thus fixing the trait in question for a particular person. With the "gene theory" the four conditions mentioned earlier are satisfactorily met and we could expect normal distributions for such traits as height, head size, nose shape, eye color, length of ear lobes, and various measures of skeletal structures. It is important to stress again the fact that no selective factors or no special external environmental conditions should be operating. For example, the distribution of weights of persons would not likely be a normal probability distribution because their weights are subject to conscious control.

Much has been said about intelligence being normally distributed. According to our previous discussion, we would expect a normal probability distribution only if intelligence is a Mendelian trait, innate and not subject to voluntary or external control. Getting at innate intelligence has been the chief stumbling block for intelligence testing and there is still a long way to go (see Chapter 14). It was mentioned earlier that the distribution of IQ's for the Stanford-Binet scale resulted approximately in a normal probability curve. This was due to a careful selection of items in the test so that the total result would be a normal curve and not because the test was getting at innate intelligence. Achievement of pupils in school subjects is also believed to be normally distributed. By now, it should be obvious that this would be likely only under the most ideal circumstances, with each pupil working to full capacity and environmental conditions fully equated.

Sampling and Error

The most important applications of the normal probability curve are in connection with sampling and error. The curve was first encountered by physical scientists in the form of random errors of measurement and to them it was known as the "curve of error" (7:166). They observed that when different persons took measurements of the same object, the results were not always the same. When several of these measurements were plotted, the resulting curve approximated what is now known as the normal probability curve.

STANDARD ERROR OF A MEASUREMENT

The standard deviation of this particular normal probability curve or "curve of error" is given a special name, the *standard error*. The standard error is often encountered in educational and psychological measurement;

therefore, let us consider a specific example to show how it is used. Suppose we have administered a standardized intelligence test to a student and determined that his IQ is 112. We could stop here and consider this to be his true IQ. However, we have an uneasy feeling that if we were to administer a different but equivalent form of the test to him, his IQ would not be exactly the same.

Now, suppose we had at hand several equivalent forms of the test that we could use to retest the student many times, what would the distribution of his IQ's be? On the basis of the experience of physical scientists, we could reasonably expect the distribution to approximate the normal probability curve. Now this is where the standard error comes in. The standard error would be the standard deviation of this theoretical distribution of IQ's obtained by retesting the student several times. If we were to look up the test manual for this particular intelligence test, we would find the standard error given. Let us say that the standard error is 4, since this is close to what is reported for most tests. We are now in a position to interpret the student's IQ of 112 with the aid of Figure 37.

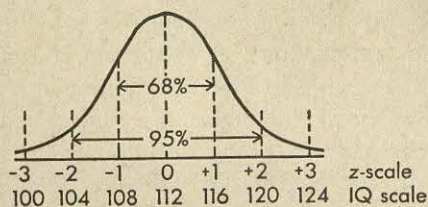


Figure 37. Standard error of an IQ illustrated.

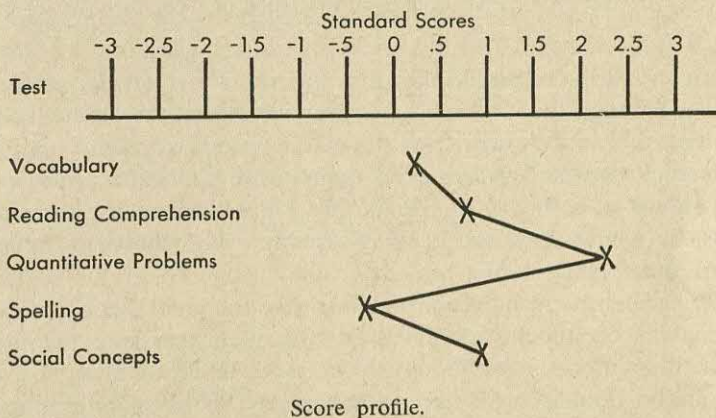
1. The chances are 68 out of 100 that the student's "true IQ" is somewhere between 108 and 116.
2. The chances are 95 out of 100 that the student's "true IQ" is somewhere between 104 and 120.

As indicated in Figure 37, the student's IQ of 112 is placed at the center of the probability curve. Subtracting and adding the standard error of 4 units to the 112 IQ gives us the interval of 108 to 116, which includes 68 per cent of the area of the probability curve. Similarly, two standard errors above and below the 112 IQ results in the interval 104 to 120, which encloses 95 per cent of the area. Both of these intervals are called "confidence intervals," and they serve as a basis for estimating true values. For the first interval, 108 to 116, we would be sure 68 per cent of the time that this interval would contain the true IQ; thus it is called the 68 per cent confidence interval. Likewise, the second interval, 104 to 120, is called the 95 per cent confidence interval.

APPLICATION OF STANDARD ERROR TO INTERPRETATION OF TEST SCORES

The standard error of a score, then, serves as a basis for establishing confidence intervals.³ Scores on all standardized tests should be interpreted in this manner and should never be thought of as fixed scores. In the physical sciences where the measurements are much more precise and accurate, the practice of providing an interval estimate is generally followed. In the test manuals of many standardized tests this very important standard error of score is provided and it should be employed more often than is generally done.

One place where the standard error of a score is particularly important is in connection with score profiles. A score profile is a graph of different test scores for an individual student that are expressed in comparable units. Suppose a student has been administered the following tests: Vocabulary, Reading comprehension, Quantitative problems, Spelling, and Social concepts. The results of these tests can be portrayed by the following score profile.



In the interpretation of this profile, we are concerned primarily with any differences to be found in the special abilities of the individual. Thus, the question of how significant are the differences in scores becomes important. And if we are to judge from the profile shown we may tend to overestimate their significance.

In Figure 38, the confidence intervals provided by the standard errors of the scores are indicated. The broad bars represent 1 standard error above and below the observed score or the 68 per cent confidence interval; and the

³ Standard errors may be computed for the mean and standard deviation also. They are called respectively the standard error of the mean and the standard error of the standard deviation. The computation of these standard errors can be found in any statistics textbook. The interpretation of these standard errors is the same as the standard error of a score. For instance, the standard error of a sample mean is used to establish confidence intervals for the true mean.

lines extend over 2 standard errors, representing the 95 per cent confidence interval.

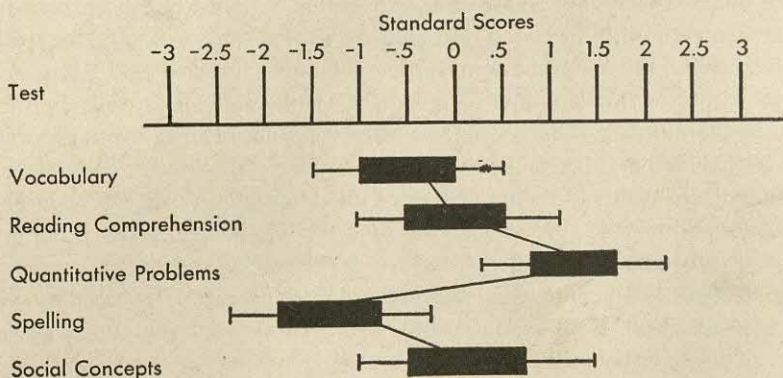


Figure 38. A test profile that includes an indication of standard errors of the scores.

The intervals in the profile shown in Figure 38 give a truer picture of the differences between the scores. Now it is apparent that the difference between the student's vocabulary score and his reading score is insignificant; whereas the difference between his score on quantitative problems and his spelling score is shown to be highly significant. We know this because the confidence intervals for vocabulary and reading overlap while those for quantitative problems and spelling do not.

Drawing such bars on each profile might be too laborious a process where large numbers of students are involved. But if the bars are not drawn, an image of these intervals should be held in mind. In educational measurement, reading an individual score is like reading a meter that has a vibrating pointer. At best, we can only establish with a certain amount of confidence an interval which we believe contains the true score.

Correlation

We turn now to another important statistical concept in educational measurement, the correlation coefficient. Fundamentally, the correlation coefficient is used to describe the extent to which the variation of one set of measurements of a variable is accompanied by the variation of a set of measurements of another variable or dimension. One of our human activities in which we are ceaselessly engaged is that of associating the variation of one variable with the variation of another variable. It is one of our primary methods of getting along in our environment. At an early age, the child is able to associate certain facial expressions of his parents with their probable behavior. He is also able to associate certain noises with certain activities that go on around

the home, and the more associations he makes, the better he is able to function effectively at home. In education we are constantly studying the correlation of different variables such as interest, attitudes, achievement, and IQ so that we may function more effectively as teachers.

Concomitant Variation. In the remaining pages of this chapter, we shall constantly refer to variations of variables as being "accompanied by" or "associated with" variations of other variables. This type of variation is often called "concomitant variation." We use such phrases deliberately in order to avoid any possible interpretation of a "cause-and-effect" relation. The fact that the variation of one variable is accompanied closely by variation of another variable does not necessarily mean that the first variable is the cause and the second variable is the effect. The correlation coefficient is not involved in causal relationships but only with concomitant variation. This will be re-emphasized more meaningfully later in the chapter.

Degree of Concomitant Variation. Our experience in associating variables has indicated that some variables are more closely associated than others. For example, we would suspect that there would be very little association between a youngster's school achievement and the number of letters in his last name. On the other hand, we would suspect that there is a closer association between a youngster's school achievement and his IQ. There are some variables that are almost perfectly associated, such as the time of day and the position of the sun in the sky. The primary function of the correlation coefficient is to indicate the degree of closeness of the association or concomitant variation of two variables or dimensions.

Direct or Positive Concomitant Variation. The other function of the correlation coefficient is to indicate whether the concomitant variation of two variables is direct or positive or inverse or negative. A direct or positive concomitant variation or correlation is the case whenever an increase in one variable is accompanied by an increase in the other variable and, likewise, whenever a decrease in the first variable is accompanied by a decrease in the second variable. An example of this type of association is depth and water pressure, because an increase in depth of water is accompanied by an increase in water pressure. Other examples of variables that have positive correlation are height and weight, IQ and school achievement, ability in mathematics and achievement in science courses, vocabulary and reading comprehension.

Inverse or Negative Concomitant Variation. Other variables are inversely associated or have a negative correlation. In this case, an increase in one variable is accompanied by a decrease in the other variable, or a decrease in one is accompanied by an increase in the other. An example of this type of variation is altitude and atmospheric pressure, an increase in altitude being accompanied by a decrease in atmospheric pressure. Likewise, in flying non-stop from Chicago to San Francisco, an increase in speed is accompanied by

a decrease in the time it takes; thus, in this case, there is a negative correlation between speed and time. Other pairs of variables having negative correlation are dominance and submissiveness, TV reception and distance from transmitter, ability to memorize digits or nonsense syllables and age after twenty years.

The Correlation Coefficient, r . On the basis of the above discussion, the correlation coefficient must indicate the degree of closeness of the correlation between two variables and the direction of correlation, whether it is direct or inverse. Generally, the letter r is used to denote the correlation coefficient and the degree of closeness is indicated by a decimal scale ranging from 0 to 1 with 0 indicating no concomitant variation whatsoever and 1 indicating perfect concomitance. Intervening decimals such as .23, .57, .82, would indicate varying degrees of concomitance with $r = .82$ indicating a closer association than $r = .57$. The direction of the correlation is indicated simply by a + or - sign, where + indicates a positive or direct correlation and - indicates a negative or inverse correlation. An $r = -.82$ would still indicate a closer association than an $r = +.57$ but the directions of the corresponding variation would be different. $r = -1.00$ would indicate a perfect inverse concomitant variation.

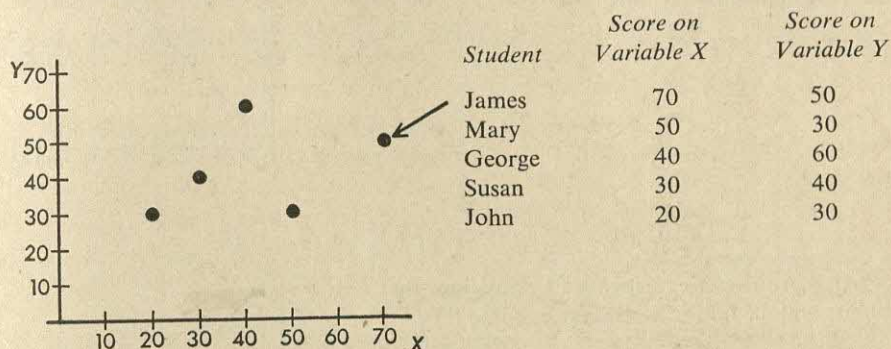
The Scatter Diagram. Before we develop a definition of the correlation coefficient, we need a visual portrayal of the relationship between two variables. This is provided by the scatter diagram.

The construction of a scatter diagram will be discussed in detail. However, the primary purpose here is to provide a figurative development of the nature of correlation and concomitant variation rather than to enable one to compute the correlation coefficient.

In order to set up a scatter diagram for two variables, we need first to have paired measures of the two variables. Suppose, for example, we wish to develop a scatter diagram for vocabulary and reading comprehension. We need not only measures of vocabulary and reading comprehension, but we also need to have these measures paired on the basis of individual students where each student has a vocabulary score and a reading comprehension score. The next step is to set up a vertical axis and a horizontal axis. On the horizontal axis we lay off a scale for one of the variables and on the vertical axis, a scale for the other variable. Sometimes it is desirable to have these scales laid off in terms of score intervals (see pages 130-133). The vertical and horizontal scales are used then in plotting the pairs of measures and the resulting scatter of points is called a scatter diagram. The steps in setting up a scatter diagram are summarized in the accompanying outline.

Outline for Setting up a Scatter Diagram for Two Variables

1. Begin with pairs of measurements for the two variables.



2. Draw a horizontal axis and a vertical axis.
3. On the horizontal axis lay off a scale for variable X and on the vertical axis lay off a scale for variable Y .
4. Plot each pair of scores as a point on the diagram. For example, for the first pair, go out to 70 on the horizontal scale, up to 50 on the vertical scale, and plot the point. Each point, therefore, represents a pair of scores for a particular student.

The scatter diagram provides us with a means of observing the concomitant variation of two variables. A few special examples are shown in Figure 39.

These scatter diagrams illustrate a few of the many possible ways in which two variables may be related. We shall be particularly interested in the last of these illustrations, in which some degree of concomitant variation has been indicated but not perfect correlation. Most of the correlations to be found between various psychological and educational variables are of this type.

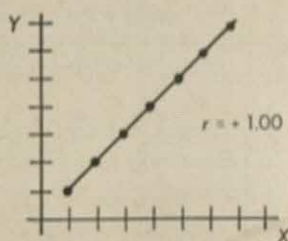
DEFINITION OF THE CORRELATION COEFFICIENT

Our concern now will be to show how the correlation coefficient indicates the degree of closeness of concomitant variation between two variables.⁴ We shall present here a definition that will be helpful in understanding the nature of a correlation coefficient rather than in computing the value of the correlation coefficient.⁵

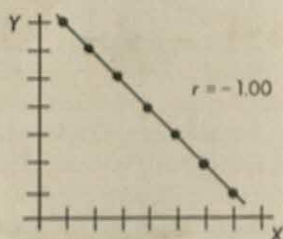
⁴ To use a correlation coefficient as an indicator of the degree of closeness of concomitant variation between two variables, it is necessary to assume that their true scatter diagram is in the form of a two-way normal distribution.

⁵ Methods for computing the correlation coefficient are to be found in any elementary statistics text. See bibliography at end of the chapter.

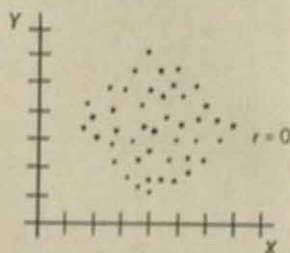
If there is perfect positive correlation between variables x and y , the points will all fall on a straight line like this:



If there is perfect negative correlation, the points will all fall on a straight line like this:



If there is no correlation between variables x and y then the scatter diagram will look something like this:



If there is some degree of positive correlation then the scatter of points will indicate a general trend as shown:

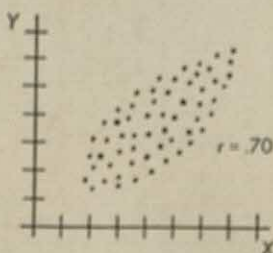
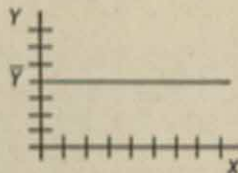


Figure 39. Illustrative scattergrams for different degrees of correlation.

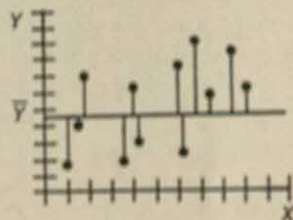
Variation Around a Line. Our definition of the correlation coefficient is based not only upon the portrayal of concomitant variation by the scatter diagram, but also upon the idea of variation of the points in the scatter diagram around certain key lines. One of these key lines is a horizontal line drawn through the mean of the Y -variable ($\bar{Y}$) in the scatter diagram, as shown:

The mean horizontal line drawn through the mean of the Y variation (Y) of a scatter diagram.



Now, suppose we draw the scatter diagram of the X and Y variables again and this time show the points in relation to the horizontal mean line. How can we determine the variation of the points around this line?

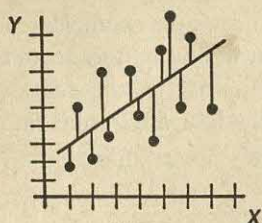
The variation of the points around the horizontal mean line is based upon the distance of each point from the line as indicated in the figure by vertical lines drawn from each point to the line. These distances are used in computing *total variance*.



If we square the vertical distance of each point from the horizontal mean line, add these squares, and divide by the number of points, the result is the *variance* (see Chapter 7) of the points around the horizontal mean line. This quantity represents the *total variation* of the Y variable. As explained in Chapter 7, if we take the square root of this variance, the result by definition is the standard deviation of the Y variable. We shall use variance here as an indicator of variation rather than the standard deviation because variances, in this case, can be added or subtracted.

Taking the same scatter of points, we notice that the scatter of points seems to have a trend as indicated by the straight line in the accompanying figure. This trend line (also called the regression line) represents the variation of the Y variable with respect to the X variable and the exact position of this line can be calculated.* Here, we shall content ourselves, however, with a line that approximates the general trend of the points, based upon our eye and judgment.

* In most elementary statistics texts, the "least squares" method for obtaining these lines is outlined or at least the necessary formulas are given.



The variation of the points around the trend line is based upon the distance of each point from the line as indicated in the figure by vertical lines drawn from each point to the line. These distances are used in computing *unexplained* or *error variance*.

Again, if we square the vertical distance of each point from the trend line, add these squares, and divide by the number of points, the result is the variance of the points around the trend line. We shall call this variance the *unexplained variance* or *error variance* since we are not able to account for, or give reasons why, the points do not closely correspond to the trend line. If we take the square root of the unexplained or error variance, the result is a standard deviation which is given a special name, "standard error of estimate."

If there is some degree of correlation or concomitant variation between the two variables under consideration, then the variance of the points around the trend line in the scatter diagram *should be less than* the total variation of the points around the horizontal mean line. In fact, the trend line by definition is that line that reduces the variance of the points to a minimum. In other words, the trend line is the best we can do to explain or account for the variation of the *Y* variable with respect to the *X* variable. This leads us to what is meant by "explained variance."

Explained variance = Total variance minus unexplained variance

This simply tells us that the explained variance is equal to the difference between the total variance around the horizontal mean line and the unexplained variance around the trend line.

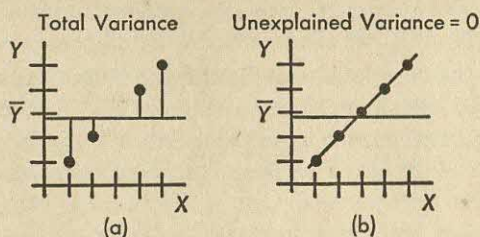
The Correlation Coefficient as a Ratio. The correlation coefficient is not different from any other coefficient, in that it is basically a ratio. The square of the correlation coefficient is equal to the ratio of the explained variance to the total variance.

$$\text{Correlation coefficient squared } (r^2) = \frac{\text{explained variance}}{\text{total variance}}$$

By taking the square root of both sides of the above equation, we can express the definition in another form.

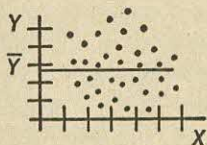
$$\text{Correlation coefficient } (r) = \sqrt{\frac{\text{explained variance}}{\text{total variance}}}$$

With this definition of the correlation coefficient, we can show how the degree of closeness of concomitant variation is indicated by considering the two extreme cases of perfect correlation and of no correlation. First we take the case of perfect correlation.



Perfect Correlation. Figure *a* presents a hypothetical scatter diagram for two variables, X and Y . The total variation of the Y variable is indicated by the distances of the points above and below the horizontal mean line shown in the figure. In Figure *b*, the trend line or regression line is drawn and all the points fall on this line. This would indicate that there is no unexplained or error variance. Thus, the explained variance and total variance are equivalent and their ratio is equal to 1. In the case of perfect positive correlation, r is equal to 1.

No Correlation. We turn now to a hypothetical case in which there is no correlation between the two variables. The scatter diagram is provided in the accompanying figure and the horizontal mean line is shown.



In this scatter of points, no trend is indicated. In fact, the trend line and the horizontal mean line are one and the same line. Therefore, the total variance of the Y variable is completely unexplained. In other words, the total variance is equal to the unexplained variance, and the explained variance is equal to zero. With the numerator in the ratio being zero, the correlation coefficient (r) then is equal to zero.

Importance of the Trend Line. We can see now how important the trend line is in analyzing the correlation coefficient. The trend line serves as a basis for dividing the total variance into explained variance and unexplained variance. The correlation coefficient is directly related to the explained variance. When the explained variance is zero, then $r = 0$, and when the explained variance approaches the value of the total variance then r approaches 1. This, then, gives us a rough graphic picture of how the correlation coefficient indicates the varying amount of concomitant variation for two variables. The trend line does not necessarily have to be a straight line. In some cases, the scatter of points may indicate a parabola or some other curve. The definition of the correlation coefficient, however, remains the same.

INTERPRETATION OF THE CORRELATION COEFFICIENT

In interpreting the correlation coefficient, the definition we have presented can be used to good advantage. Let us suppose that a certain study reports that the correlation between two variables is .80. Using the definition, we find that

$$r^2 = .64 = \frac{\text{explained variance}}{\text{total variance}}$$

By squaring the correlation coefficient $r = .80$, we find that the ratio of explained variance to the total variance is .64. In other words, 64 per cent of the variance of the Y variable is accounted for by its concomitant variation with the X variable. Likewise, if $r = .40$, then $r^2 = .16$ and 16 per cent of the variance of the Y variable can be attributed to its covariation with the X variable. On the basis of being able to explain variance, we can say that an $r = .80$ indicates a degree of correlation 4 times as great as an $r = .40$. Furthermore, the definition can be used to show that the difference in correlation between $r = .80$ and $r = .90$ is much greater than the difference in correlation between $r = .20$ and $r = .30$. Therefore, the increase in the amount of covariation as indicated by the correlation coefficient is not evenly spread out from $r = 0$ to $r = 1$. Nor does $r = .50$ indicate a halfway point in concomitant variation, since it represents only 25 per cent covariance.⁷

Correlation Does Not Mean Cause. A final word needs to be said about the fact that a high correlation between two variables does not necessarily mean that variation in one variable will *cause* a certain amount of variation in the other variable. From our development of the correlation coefficient, it should be apparent now that covariation is the sole concern. Whenever we find correlation existing between two variables, all we can say is that the variation of one variable "is accompanied by" a certain amount of variation in the other variable, depending on how high is the correlation coefficient. It is quite possible that the correlation found between two variables is entirely coincidental. Studies have shown, for instance, that there is a positive correlation between the number of inmates in insane asylums and enrollment in colleges. Also, a positive correlation between the salaries of Presbyterian ministers in Massachusetts and the price of rum in Havana was found over a

⁷ Often, instead of comparing ratios of variances, a comparison of standard deviations is used in interpreting the correlation coefficient. Specifically, the standard error of estimate based on the trend line is compared to the standard deviation of the Y variable based on the horizontal mean line. The ratio of these two standard deviations is called an index of predictive efficiency for the correlation coefficient. Tables and curves are provided in most elementary statistics texts. However the essential characteristics of the correlation coefficient, which are mentioned here, hold true whether we compare variance or standard deviations.

ten-year period.⁸ In these instances we would strongly suspect that the correlation was purely coincidental and we would not be likely to attach any significant causal relationship to the results.

It is also possible that the correlation found between two variables was due to their both being related to a common factor. There is generally a high correlation between vocabulary and reading comprehension, but we cannot say which is the cause of the other. In fact, both of these variables may be due to the cultural level of the home environment. The correlation coefficient, then, should not be used to establish a cause-effect relationship between two variables.

Reliability Estimates

In Chapter 3, reliability was mentioned as one of the essential characteristics of a good measuring procedure. Briefly, in connection with reliability such questions are raised as: How accurate is the measuring procedure? How reliable? How precise? How consistent? How trustworthy? How much confidence can we place in the result? With an understanding of the normal probability curve and the correlation coefficient, we can now investigate more technically and analytically the meaning of reliability, and the types of data and statistical procedures necessary for estimating the reliability of a measuring procedure.

For our purposes, to say that reliability is synonymous with dependability, accuracy, and consistency is insufficient. We need an operational definition, a definition that provides us with a method for determining reliability. In casting about for an operational definition of reliability, we might ask ourselves the question: How would we ordinarily check the reliability of a person? We could take a statement or observation made by the person and have the information checked independently by other observers. If there was close agreement between the original statement of the person and the report of the observers, then we would have evidence that the person is reliable. In science, the reliability of an experiment is checked in much the same fashion. The experiment is conducted independently by several scientists and the results are checked for closeness of agreement with the original experiment. Likewise, a machine is said to be reliable when its successive performances are observed to be approximately identical.

From these few examples just presented, we can see that the process of determining the reliability of an observation or procedure is essentially a matter of comparing one or more independent observations or performances with the original observation or performance and checking for closeness of agreement. This concept of reliability provides us now with an operational basis for determining the reliability of an educational measuring procedure. We simply obtain independent observations of the results of the measuring procedure and then determine how closely these observations agree. The prob-

⁸ See Harold Larrabee, *Reliable Knowledge* (Boston: Houghton Mifflin Co., 1945).

lem of determining how closely the observations correspond is essentially a statistical problem and will be discussed after an analysis has been made of the possible sources of variation among observations.

In educational measurement, we are concerned primarily with human responses and, therefore, the sources of inaccuracies of measurement are necessarily complex. It is doubtful, then, if one can develop an all-inclusive list of sources of variation in educational measurement. However, the following are among the more important ones.

POSSIBLE SOURCES OF VARIATION IN HUMAN RESPONSES TO EDUCATIONAL MEASURING PROCEDURES

1. *Sampling Variation.* In educational measurement, the measuring procedure usually consists of a sample of questions, tasks, or observations taken from an exceedingly large number of possible questions, tasks, or observations. Human responses vary with different samplings of items and, hence, this is a source of unreliability in a measuring procedure.

2. *Psychological and Physiological Variation.* While an individual is undergoing measurement, he is subject to certain physiological and psychological determinants that may cause variations in his responses. Some of these factors are health, fatigue, motivation, tenseness, distraction proneness, tendency toward guessing, and memory fluctuations. Also included in this category are such environmental factors as heating, lighting, noise, interruptions, and seating facilities, which may affect the individual psychologically and physically.

3. *Variation in the Technical Aspects of the Measuring Procedure.* A third source of inaccuracy in educational measurement is in the quality of the measuring instrument itself. This category includes variation in the adequacy of directions, in clarity of test items, in typographic quality, and in the scoring or rating procedure.

These sources of inaccuracies all have one common characteristic. The nature of their influence in any given measurement is somewhat a function of chance. During any measurement procedure, there is some chance of a certain factor influencing the result unfavorably but perhaps an equal chance that it will increase a measure or improve a score. There is also the possibility that it will not affect the measurement at all. Consequently, as we have observed earlier (Chapters 3, 4, 5 & 6), increasing the length of a measuring procedure or repeated measurement will permit chance to operate as much to increase scores as to depress them and thus increases the reliability of measurement.

It should be noted that psychological and physiological factors are likely to fluctuate at different rates. Some of these factors are relatively constant during a period of measurement and even for several days or weeks, for example, health, motivation and proneness to distraction. Others, such as fatigue, noise and interruptions, are operative perhaps for only a minute or at

most for the period of the given measurement. These varying rates of fluctuation tend to complicate the matter of repeated measurement. If measurement is reapplied after a short interval, the factors which fluctuate rapidly may be taken care of but those which fluctuate slowly may have remained unchanged and may continue to affect the measurement in the same way. If done after a long enough interval for these long term factors to fluctuate normally, then the individuals being measured may have undergone substantial changes with regard to whatever is being measured.

DETERMINING THE RELIABILITY OF MEASURING INSTRUMENTS⁹

We turn now to the procedures used to determine the reliability of measuring instruments employed in education. According to our previous statements about reliability, we should have several independent observations of the results of the given instrument in order to check for closeness of agreement. In measuring human behavior, though, we know that the individual may be changed by the measuring process. For example, in taking a test, a student may benefit by the practice and thereby increase in "test-wiseness" and even in knowledge of the material. So it is necessary to keep the number of repeated measurements to a minimum and even to gauge reliability in other ways than through repeated measurement.

The three procedures generally used for determining the reliability of educational measuring devices and processes are:

1. *Retest*—repetition of the same measuring procedure.
2. *Subdivided Test*—where for purposes of scoring a test is subdivided into two or more parts for comparison.
3. *Equivalent Test*—where a parallel test is developed and administered for comparison with the original test.

Retest Procedures. Reapplication of the measuring device or procedure would seem to be the most straightforward approach to checking its reliability. For instance, if we wish to check the reliability of an instrument that measures a person's strength of grip, we would simply take repeated measurements for the same individuals and check the results for consistency. The same procedure would be applicable to all instruments measuring such motor abilities and physical characteristics as vision, hearing, weight, ball-throwing, and reaction speed. However, when we consider tests that measure psychological characteristics, we have noted that this procedure of repeated measurement runs into difficulty. In addition to the practice effects already mentioned, the individual's second approach to the same test will differ from his first approach since the novelty has worn off, the directions are already understood, and only a cursory glance is needed for the reading passages and questions. Another limitation of the retest procedure is the fact that there is no provision for

⁹ While reliability estimates usually are confined to tests and the discussion here is in terms of tests, the procedures described are applicable with appropriate modifications to observation techniques, to rating scales and to the scoring of products as well.

varied sampling. The test items remain the same for each performance. Chance fluctuations in physiological and psychological conditions may be provided for by this method if the interval between tests is properly chosen. Moreover, the method is well adapted to appraising variations in technical aspects of the test.

Subdivided Test Procedures. The second procedure for determining reliability, the subdivided test, represents an attempt to minimize the practice effect encountered in the retest method. Generally a test is administered to a group and then it is subdivided in some fashion for comparison of results. The test, after it has been administered, usually is split into equivalent halves by putting the odd-numbered questions into one of the halves and the even-numbered questions into the other. The two halves are then scored separately and are compared for consistency. This procedure is generally called the "split-halves" method for determining reliability and it is the one most widely used because of the convenience of providing an estimate of reliability from a single administration of the test.

Although the "split-halves" method reduces the practice effect to a minimum, it does have some important limitations. In the first place, this procedure does not provide for the day-to-day or week-to-week fluctuations of human physiological and psychological factors. These factors may not have an opportunity to vary during a single administration of a test. Unreliability in the technical aspects of a test and sampling variation are adequately appraised by this procedure only if the test is sufficiently long.

Equivalent Forms Procedures. The third procedure for determining the reliability of a test is to develop another test according to the same specifications but with different questions, and then to compare the results of the two equivalent tests. If there is a time lag of a few weeks in the administration of the two parallel tests, then each of the three categories of variation previously mentioned has an opportunity to affect the testing situation and practice effect is minimized. This equivalent test procedure then provides the most severe estimate of reliability and, for this reason, the procedure is preferable to the other two for estimating the reliability of tests. There are, however, certain practical limitations to this procedure. The problem of constructing an equivalent test is a great one and the administration of a second separate test to the same individuals may place an unreasonable demand upon the individuals. The short cuts offered by the other two procedures are tempting, but they are taken at the expense of accuracy in estimating the true reliability of the test.

STATISTICAL DETERMINATIONS OF RELIABILITY

Having considered the various factors involved and the methods for determining the reliability of educational measuring procedures, we are ready now to outline briefly the statistics used in determining the degree of reliability. As was pointed out in the definition of reliability earlier in this section, the

problem of determining how closely successive observations of a measuring procedure correspond is essentially a statistical problem.

Reliability Coefficient. There are essentially two ways of expressing statistically the reliability of an educational measuring procedure. One way is to indicate how well each individual's score on the first testing or on the one half of the test corresponds to his score on the second testing or the other half. The degree of correspondence between the two results may be expressed statistically as a correlation coefficient. A high correlation coefficient indicates a close correspondence between the individual's scores in the two measurements and a high degree of reliability for the measuring procedure. This correlation coefficient is more generally known as the reliability coefficient and is designated by the symbol r_{11} (read as "r sub one-one").

Standard Error of Measurement. The second method of expressing reliability statistically is to indicate the amount of variation in the repeated measures of an individual. This is done by using the standard error concept discussed earlier in this chapter. Owing to the somewhat random nature of effect of the error factors involved, the repeated measurements of an individual tend to form a normal probability curve. The amount of variation for these repeated measures is given by the standard deviation of this probability curve and is known as the *standard error of measurement*. In educational measurement, we seldom obtain more than two observations of a measuring procedure for each individual, but from these two observations it is statistically possible to estimate the variation of a larger number of repeated observations. Thus, a smaller standard error of measurement indicates a higher degree of reliability for the measuring procedure.

The two statistical measures of reliability, namely, the reliability coefficient and the standard error of measure, are related in the following fashion:

$$SE = S \sqrt{1 - r_{11}} \quad \begin{array}{l} SE = \text{standard error of measurement} \\ S = \text{standard deviation for the test} \\ r_{11} = \text{reliability coefficient} \end{array}$$

INTERPRETING RELIABILITY DATA

We might now briefly consider the general problem of interpreting reliability data. For instance, suppose it is reported that a certain standardized test has a reliability coefficient of .89. How shall we interpret it? One way is to find out and to compare the reliability coefficients for other similar standardized tests. This procedure makes the interpretation of reliability entirely a relative matter. Another way is to determine the standard error of measurement for the test and then decide whether or not the results of the test are sufficiently accurate for the desired purpose.

In interpreting reliability data and in comparing the reliability of different tests, certain general considerations should be kept in mind.

1. *How accurate results are needed?* If the purpose for using the test is

to determine the relative standing of an individual in a fairly homogeneous group, then a very reliable test is necessary. On the other hand, if the purpose is to determine the general grade level of achievement of the individual, then a less reliable test may be used.

2. *What is the range of the group being measured?* In general, the test measuring a group that presents a greater range in the trait being measured will have a higher reliability coefficient than a test measuring the same trait for a more homogeneous group. This fact should be remembered when comparing the reliability of two tests.

3. *Is the test too easy or too difficult for the group being measured?* In general, for a test that is too difficult, an undue amount of guessing may result, thereby lowering the reliability of the test. When a test is too easy for a group, it will fail to discriminate among the members of the group and again its reliability will be lowered.

4. *What is the nature of the measurement symbol resulting from the test?* The fineness or grossness of the scaling, ranking, or classifying symbol has an effect on the reliability of the test. In general, the measuring procedures using scale symbols are more reliable than measures using classification symbols. Essay tests that use classifying symbols generally have reliability difficulties.

Validity Estimates

As previously defined, validity has to do with aim—the capacity of a measuring procedure to measure what it purports to measure. There are essentially two approaches to validity. One is called the *logical or rational* approach, in which the validity of the test is simply analyzed in terms of its objectives, the character of its items, and its general format. This approach has been amply discussed in Chapter 3.

The approach we are concerned about here is called the *empirical or statistical* approach to validity. A specific example probably will show best how a test is validated by this method. Suppose, for instance, we wish to check statistically the validity of a mechanical comprehension test. We would first select a typical group of persons for which the test is intended and administer this test to them. Then we would have each individual in the group be observed and rated by a group of experts in actual situations that demand varying levels of mechanical comprehension. A correlation coefficient is computed between the test scores and the ratings of the experts and a high correlation would be construed to indicate that the test has high validity. The ratings of the experts is called the *criterion* and the correlation coefficient computed for the test is called the *validity coefficient*.

As stated above, the empirical approach to validity first requires the establishment of a criterion upon which the test is validated. This criterion should provide a reliable measure and be as free from bias as possible. For some kinds of tests it has not been possible to establish such a criterion. Achievement tests are an example. For aptitude tests or tests used for prediction, it has been

generally possible to establish a criterion in terms of later performance. Some tests are validated with reference to a similar test for which validity is assumed. For example, some intelligence tests have been validated on the basis of their correlation coefficients with the Stanford-Binet intelligence test. Since the setting up of a suitable criterion presents a difficult problem, the statistical approach to validity is somewhat limited.

Summary

The normal probability curve, the basis for many analyses and interpretations of measures, is a theoretical distribution of expected probabilities for any given occurrence when chance alone determines occurrence or nonoccurrence and the opportunities for occurrence are unlimited. The special property of the curve making it useful in educational measurement is the predictable relationship between distances along the base line from the mean of the distribution and areas under the curve. The latter may be interpreted as given proportions of the distribution. The curve, then, provides a standard way of estimating the operation of chance in behavioral measurement and for predicting the distribution of genetically determined human traits or dimensions.

The *standard error* of measurement is the standard deviation of the normal distribution of any measure that is repeated an unlimited number of times. Thus it is an index of scatter or chance variation in the measure. If the standard error of a pupil's score on an achievement test were 5 and his score were 80, the odds are 2 to 1 that his true score lies someplace between 75 and 85.

In education, concomitant variation between two sets of measurements is an important consideration. The correlation coefficient is a mathematical expression for the extent to which variation in one set of measurements of a variable is accompanied by variation in a set of measurements of another variable. The coefficient is symbolized by r and has magnitude from +1 through zero to -1. As r approaches +1, the relationship between the two sets of variables is direct or positive and extensive. As it approaches -1, the relationship is inverse or negative but equally extensive. The coefficient itself indicates association only. Any cause-and-effect relationship is a matter of inference.

Coefficients of correlation are used as indexes of the estimated reliability of tests and other measuring instruments. Reliability means the extent to which a measuring procedure is likely to yield the same measure when it is reapplied to the same dimension of the same phenomenon. The reliability of tests may be assessed directly in some cases by actually reapplying the test to the same group and then determining the extent of correlation between scores on the first administration and on the second. Generally, though, it is estimated indirectly by comparing the scores from comparable forms of the test or the scores from two equivalent halves of the test.

Validity of a measuring procedure (how well it measures what it purports to measure) may be expressed by a coefficient of correlation when there is a

companion procedure accepted as a criterion or standard and with whose results the results of the procedure in question can be compared. Usually, though, there is no accepted criterion and hence statements about validity must be based on a logical analysis of the procedure and the measures it obtains for persons having certain known characteristics.

EXERCISES

1. Why is the term "normal curve" a misleading term?
2. A student obtained a score of 84 on a test for which the standard error is 3. What is the 68 per cent confidence interval? the 95 per cent confidence interval?
3. Give examples of some possible uses of the normal probability curve to the classroom teacher.
4. What assumptions must be made by a classroom teacher who "grades according to the curve"?
5. Suppose that the correlation between vocabulary and reading comprehension is .80. What proportion of the variance in reading comprehension is due to variation in vocabulary?
6. The scores on a spelling test are normally distributed with a mean of 65 and standard deviation of 15. What is the probability that a student's score will be greater than 80?
7. Examine the test manuals of three different standardized tests in regard to what is said about the reliability and validity of the tests. Make a critical analysis of each in the light of what you have learned in this chapter.

BIBLIOGRAPHY

1. Adams, J. K., *Basic Statistical Concepts*. New York: McGraw-Hill Book Co., Inc., 1955.
2. Clark, Charles E., *An Introduction to Statistics*. New York: John Wiley & Sons, Inc., 1953.
3. Croxton, F. E., and Cowden, D. J., *Applied General Statistics*. New York: Prentice-Hall, Inc., 1940.
4. Freund, John E., *Modern Elementary Statistics*. New York: Prentice-Hall, Inc., 1952.
5. Guilford, J. P., *Fundamental Statistics in Psychology and Education*, 3d ed. New York: McGraw-Hill Book Co., Inc., 1956.
6. Johnson, P. O., *Statistical Methods in Research*. New York: Prentice-Hall, Inc., 1949.
7. Patterson, C. H., "A Note on the Standard Error of Measurement," *Journal of Educational Psychology*, 46:239-242, April 1955.
8. Snodgrass, Florence T., "Unreliability of Group Test Profiles," *Journal of Educational Psychology*, 45:129-142, March 1954.
9. Sprowls, R. C., *Elementary Statistics for Students of Social Science and Business*. New York: McGraw-Hill Book Co., Inc., 1955.

10. Walker, Helen W., *Elementary Statistical Methods*. New York: Henry Holt and Company, Inc., 1943.
11. Wilks, S. S., *Elementary Statistical Analysis*. Princeton, N. J.: Princeton University Press, 1951.

CHAPTER 9

EVALUATIVE STANDARDS—MARKING AND REPORTING ACHIEVEMENT

In the introduction to this first section of the book we defined measurement and evaluation as being a twofold action. In the first phase, the status of a phenomenon was appraised in some manner and in the second, the value of this status was determined by comparing it with an appropriate standard. So far we have dealt exclusively with the first phase, measurement itself. We have examined the varied procedures by which appropriate measurement symbols may be assigned to measurable dimensions of educational phenomena. For the most part these symbols have been numerals denoting classification or rank and the procedures have been observation, product analysis, and several types of testing. Now we wish to discuss the second phase, evaluation, and its application to schooling. In education, and more particularly in the teaching situation, measurement symbols seldom are used except as a basis for making qualitative judgments about the achievement of pupils. Johnny's test score usually is not left to stand by itself but is the basis for a declaration that Johnny's learning is excellent, good, fair, poor, or unsatisfactory, and the making of such statements seems to be one of the major functions of a teacher.

Distinction Between Evaluation and Measurement

Evaluation is a process of making qualitative determinations. As such it is akin to what we have called measurement, if not just a special form of measurement. Just as we assign symbols to phenomena to describe their status, so do we assign symbols to phenomena to indicate their quality or desirability. And as we often use a scale of inches, pounds, or seconds as the basis for measuring status, so do we often use a scale of quality variation as the basis for evaluating the phenomena. We call these quality scales *evaluative standards*. To show the distinction between measurement and evaluation as we are using the terms, consider the following illustration. Imagine that we wish to buy some lengths of lumber for a house we are building. The evaluative standard to be applied to the lumber is the blueprint for the house. We go to a lumberyard and measure with a tape measure the linear dimensions of several pieces. When we compare these measurements with the lengths called for by our evaluative standard (the blueprint), we find that certain pieces

would fit but that others would not, and so we qualitatively rate the pieces of lumber as being desirable or undesirable for our purpose. In this example, the measurement function and evaluation function are clear-cut. It was the function of measurement to provide the symbols representing the status of the pieces of lumber at that time. It was the function of evaluation to use the evaluative standards in making a qualitative judgment.

From the example it may seem that the distinction between evaluation and measurement will always be clearly discerned. On the contrary, there are times when the difference between measurement and evaluation is subtle and difficult to identify. This occurs when evaluation becomes automatic without any conscious thought after the measurement has been made. A common illustration of this is when the relative size of a test score automatically connotes relative standing with regard to desirability; where it already is understood that the top score represents top quality and excellence, and so on down to where low test scores automatically represent unsatisfactory standing. Another example of the unclear distinction between measurement and evaluation is the case where custom and long usage have fixed an association between certain measurement symbols and particular standards of quality. For instance, certain IQ brackets through long association have come to represent certain "values" of intelligence: 20-50, "imbecile," 50-70, "moron," 140 and up, "genius." Because of the prevalence of this automatic translation of measurement symbols into evaluative symbols it must be emphasized that measurement symbols are not in themselves standards and it is only by custom that they may directly symbolize value.

Chapter 4 presents still further examples of confusion between measurement symbols are not in themselves standards and it is only by custom that and observations that measure or appraise status are often confused. We are more prone to observe that Jimmy is a natural-born trouble-maker than to observe objectively what Jimmy actually does. This points to the serious consequences of confusing measurement and evaluation. Measurement carries with it a certain air of objectivity and finality. Thus when a person thinks he has measured but actually has evaluated, his evaluation unfortunately takes on this air of finality and objectivity. Jimmy has been "measured" and found to be a natural-born trouble-maker. His status has been appraised and there is nothing more that can be done about it.

More examples of the hazy distinction between measurement and evaluation could be provided. However, the too prevalent misunderstanding about what is measurement and what is evaluation, and the possible consequences of this misunderstanding in the form of arbitrary action and injustice, should be sufficient to indicate the need for a full discussion of the evaluation process apart from the measurement process. We shall begin with an investigation of the nature and source of evaluative standards and symbols and then we shall discuss the evaluation process as a whole. After this we shall study briefly the

matter of marks and reporting systems. Our treatment shall for the most part be restricted to evaluating achievement in school subjects.

Nature and Source of Evaluative Standards

In a sense, an *evaluative standard* is anything that is used as a basis for judging value or desirability. In our lumber illustration the blueprint was an evaluative standard. Sometimes a *purpose* is a standard. If pupils are trying to paint a mural, actions may be evaluated in terms of their contribution to the completion of the mural. Frequently, evaluative standards are purely *arbitrary conceptions*, e.g., 70 per cent is passing, 95 per cent is excellent, above average is good, etc. In such matters as dress, talk, manners, and letter form, the standard for evaluating merit usually is *custom*—what the greatest number of a given group do. As we shall see later, standards differ as to their validity and appropriateness.

One type of standard has particular significance for school and it will be discussed at length in a later section. This is a scale or hierarchy of performance. It may consist of gradations of handwriting deemed to cover the range of possible quality in school children. It may be the typing speeds that must be attained if certain marks are to be received. Or it may be gradations in understanding of a subject that have been defined and set down.

All too frequently we find in schools that evaluation standards are based only upon emotions and rigid preconceived ideas. How I, the teacher, just happen to feel when I think of Johnny is one of these and, unfortunately, too often this one is the primary basis for the mark Johnny receives. Another is the body of prejudice stereotypes about good and bad pupils, national and racial "types," and the "athlete," the "student," and the "clown." Evaluations based upon such irrational and emotional standards are more properly termed prejudices or opinions. Needless to say, the use of irrational standards tends to vitiate effective evaluation.

The ultimate source of all evaluative standards is, of course, the value complex of our American culture. Their immediate source is the writings of experts in philosophy, psychology, sociology, history, and in other social sciences who have studied this complex. By observing what people say is good or bad, by observing the choices they make, where they spend their time and money, by examining legal documents, religious documents, and other expressions of ethics, and by looking at customs and traditional practices, these experts are able to express a consensus of what culture believes to be worthwhile.

Many of the standards used in our schools are derived directly from the general values of our culture. Thus, punctuality is an American value, so standards of pupil citizenship place a premium on being on time and a stigma on being tardy. Knowledge of our country's past being prized by American citizens, standards of knowledge for school pupils at all levels are heavily weighted in the direction of the facts of United States history. Nowhere is the relationship between culture values and school standards more apparent than

in the case of social adjustment. As a group, Americans esteem extroversion and varied pursuits. So in the schools, many guidance officials are concerned about taciturn children and those who do only one thing, and tend to label the leaders and the active as well adjusted.

Other standards in school use are particular just to schools themselves. They derive principally from the practice of countless teachers, the nature of subjects taught, and the developmental characteristic of pupils. An example of the first is the standard that "70 per cent is passing." The second source is illustrated by the premium placed on accuracy in arithmetic. The influence of child development is seen in the differential standards held at successive grades for such subjects as reading, writing, art, and music.

CRITERIA OF VALID STANDARDS

The selection or development of evaluative standards is a crucial undertaking. Not only do the evaluative standards reflect the objectives of a school program but also, as we shall see later, they have an important effect upon the construction of measuring devices. As a guide, therefore, to identifying and devising valid evaluative standards, the following criteria are presented:

1. *The evaluative standards should consist of a variation scheme of quality, desirability, or value.* There should be no question about the fact that the standards are concerned with quality or value only. This means, then, that qualitative terms should be explicitly expressed and not vaguely implied. Also, symbols that tend to confuse variations in quality with variations in status should be avoided.

2. *The degrees of variation of quality should be clear and well defined.* There should be distinct boundaries between the different categories of value and the meaning of each category should be explicit.

3. *The evaluative standards should be reasonably stable and objective.* That is to say, the standards should not be subject to sudden arbitrary changes, nor should the standards be expressed in vague terms that will permit capricious interpretation. Pupils should feel that the standards applied to them will provide consistent and fair evaluations of their performance.

4. *The evaluative standards should be expressed in terms that are appropriate and favorable to the best procedures of measurement.* Too often we encounter standards stated in such a way that it is well-nigh impossible to set up a measuring procedure to provide a satisfactory basis for making an evaluation. Suppose you were asked to make a qualitative judgment on someone's "ability to think effectively" or on someone's "personal satisfaction in achievement." You might agree that these terms do *not* facilitate measurement. Yet such general phrases often are the elements of evaluation schemes. We shall find that evaluative standards and measuring procedures are highly interdependent in the evaluation process. Properly stated standards generally provide for the best measuring procedures and likewise sound measuring procedures provide the best basis for developing adequate evaluative standards.

5. *The evaluative standards should be consistent with scientific findings about learning and child development.* The standards should be neither too high nor too low for the maturity of children in given grades. Nor should they be inconsistent with the way we learn. An example of the latter would be the placing of a high qualitative rating on a student saying or writing the right words without understanding their meaning. Another example would be the placing of a high qualitative rating on a correct numerical answer to an arithmetical problem without determining if the process was understood by the student. Other psychological aspects of evaluative standards will be discussed more fully later.

6. *Finally, the evaluative standards should be consistent with the values of our culture.* For one thing, they should reflect the emphasis in our culture on human worth and dignity, and on democratic values in general. A low qualitative rating, for example, should be placed on docile or completely submissive behavior. For another, they should relate realistically to standards of work and production held by business and industry. Without this, pupils are apt to be unrealistic in their first attempts at vocational adjustment.

A CRITIQUE OF SOME EVALUATIVE STANDARDS IN CURRENT USE

Now that we have before us some criteria for judging the validity of evaluative standards, let us examine three types of standards sometimes used in the schools.

1. Standards based upon arbitrary portions or points of distributions.
2. Standards based upon arbitrary amounts of work completed or percentage of questions answered correctly.
3. Standards based upon emotions and preconceived stereotypes.

Distribution Standards. An example of the first type of standard is provided by a teacher who will assign only 10 per cent *A*'s, 25 per cent *B*'s, 40 per cent *C*'s, and so on, to any distribution of test scores. Ordinarily these percentages are applied to distributions of whatever groups are being dealt with by the teacher at the time. Standards of this type fail to meet the first criterion listed in the previous section. No variation in quality is explicitly stated. Quality of performance has been confused with relative standing in arbitrary groups. Since groups vary greatly in performance, this type of standard fails to meet the criterion of stability. An excellent student may be penalized unduly for being in a superior group. The same criticism applies to evaluative standards based upon certain points in the distribution, such as the average (mean or median), and also to standards that claim to be based upon portions of the normal probability curve ("grading on the curve").

Percentage Standards. The second type of standard is used by a teacher who assigns an *A* to students answering at least 95 per cent of the questions correctly, a *B* for 85 per cent answered correctly, and so on; or by a teacher who assigns grades on the basis of the number of projects or experiments com-

pleted. Again, this type of standard may fail to meet the first criterion. Variation in quality is only implied and may not be properly related to variation in performance. Likewise, since tests do vary in over-all difficulty, the stability of this type of standard is open to question. For example, a superior student may be penalized by an unduly difficult test and a poor one favored by an easy test.

Emotional Standards. The third type of standard obviously fails to meet the criteria because of its extreme instability and capriciousness. As soon as emotions and feelings enter into standards, all objectivity goes out the window. Teachers will always be subject to temporary feelings of fatigue, anger, and apathy that will color their evaluations. Also, there are likely to be present unconscious stereotype reactions and unconscious refusals to reconsider values.

While these "standards" do not constitute valid bases for evaluation, many others used by teachers do. The nature and development of these are the burden of this chapter.

Evaluative Symbols

In the fourth criterion of essential attributes of valid standards, it is implied that they may be expressed in various ways, some of which are inappropriate for sound measurement. This raises a question as to what are the symbols or forms used in expressing evaluations. This question sounds familiar and, in fact, is like the one we asked about measurement in Chapter 1. We have already indicated that evaluation is essentially a process of measuring quality. Therefore it is reasonable to assume that the symbols of evaluation can be classified in the same manner as the symbols of measurement, i.e., symbols that indicate a scale position, symbols that indicate rank or order position, and symbols that classify or describe.

In Chapter 1, it is stated that a *unit* of measure is required for symbols that indicate scale position. Consequently, in order for evaluations to be expressed in the form of scale symbols, a unit needs to be established. So far, and excluding certain research situations, no unit of quality or desirability has been established, "quality" being such an ambiguous consideration. Therefore, we shall not expect to find evaluations expressed in the form of scale symbols.

We should, though, expect to find and do find in education that evaluative standards and qualitative judgments utilize rank symbols and classification symbols. Ordinarily the letters *A, B, C, D, F* are used to indicate the quality of classroom performance. In other places the words "excellent," "good," "fair," "poor" may be used, or possibly numbers such as 3, 2, 1, 0, -1 are employed. All these symbols are classificatory since each one represents a category of quality. Also we sometimes encounter percentile ranks and other rank or order symbols, which imply standing in regard to quality.

In using any scheme of evaluative symbols, it is essential that they be related to unambiguous and appropriate statements about the quality grada-

tions they represent. When the symbols stand alone, they have very limited meaning. Suppose, for example, we were informed that a sixth-grade student was given an *A* in spelling. The basis for the qualitative rating, the *A*, is not apparent. We know only that he was given a top rating. It might be generally assumed that the rating was made simply on the basis of his being able to spell a large number of words correctly. However, we must wonder about the words this student was able to spell correctly while others failed to do so. Were these words phonetically obscure? How often have these words been previously encountered? Moreover, we should wonder *how much* difference in spelling ability exists between our pupil who received an *A* and another who was given a *B* or a *D*.

The full role of standards in the evaluation process should be apparent by now. The evaluative standards are the point of reference for qualitative ratings; and, after these ratings have been made, the evaluative standards are the basis for interpreting the evaluative symbols that have been assigned. As was indicated in the previous paragraph, if the evaluation symbols are presented without knowledge of the standards to which they refer, the meaning of these symbols is limited to an interpretation of relative standing only.

Too often in education, though, just the results of evaluation, the assigned symbols, are presented. The evaluative standards used in the process are at best only vaguely implied in the test or the instrument used. This means that the person whose performance is being rated must take on faith the fact that the evaluative standards used are valid. The same faith is expected of the parents and friends of the individual being rated, and is even required of other educational institutions.

Steps in Valid Evaluation

Now that we have examined the nature of standards and the use of evaluative symbols, we wish to discuss in detail just how a teacher may go about evaluating the achievement of pupils. In doing this we shall present first an "example" and from this develop certain steps in valid evaluation and an example of a defensible standard. Assume that Miss *X*'s class has just completed a topic and she now wishes to evaluate her students' "comprehension of the material covered." Following custom, she begins to prepare a test. As she makes up questions to ask on the test, her first concern is coverage. She wants to be sure that the questions on the test cover all or at least most of the important points discussed in class. Ordinarily, this consists of thumbing through the text or some written material that was used and asking a question or two from each page. Her next concern is to make certain that her test includes some easy questions, which most of her class will answer correctly, and also some more difficult questions which only a few can answer correctly. These questions are all assembled into a test and administered to her class. After the tests are scored, she assigns qualitative ratings, *A*, *B*, *C*, *D*, *F* on some arbitrary basis, making sure that the number of *A*'s, *B*'s, *C*'s, etc., assigned corre-

sponds roughly to some preconceived distribution. Thus the evaluation process is ended.

Now that we have before us an example of what is often done in the name of evaluation, let us analyze how evaluation may need to be performed if it is to be valid. As we proceed, we shall refer frequently to what was done or not done by Miss X.

1. *Determine what is to be evaluated.* The first step in the evaluation process seems to be obvious. It should consist of stating that which is to be evaluated. What may be evaluated in this instance is no more or less than some measurable dimensions of pupil achievement. In Chapter 2 we established certain conditions of measurability and these again are applicable in establishing the dimensions to be evaluated. The three most immediately relevant conditions are that the dimensions have observable variations, be clearly defined, and conducive to agreement among observers. These conditions necessitate that the learning or achievement or "pupil objectives" to be evaluated *must* be stated in terms of behavior or performance. For example, an objective, "knows a valid argument," would be inferior to the objective, "is able to identify an unstated assumption necessary for arriving at a certain conclusion." The latter, being a behavior, is more observable, better defined, and in other ways comes closer to meeting the criteria set forth in Chapter 2. In Miss X's case the focus of evaluation seems to be something rather vague, not directly observable, and hence difficult to evaluate—"comprehension of the material just completed."

2. *Establish a standard.* After determining what we wish to evaluate, the next step usually is to select or devise a standard appropriate to its evaluation. While we have stated that several different types of standard are used for evaluating pupil achievement, the one that seems to be most valid is a performance scale or hierarchy. Moreover, if our focus of evaluation is to be performance or behavior, a standard set in the same terms is indispensable. Our standard should consist of descriptions of several levels of performance relative to each learning objective or dimension of achievement. These should range from the least "valuable" performance to the most "valuable" and should be stated in clear behavioral terms. (An example is shown on page 202.) This step of establishing a standard is only vaguely implied by Miss X in her selection of easy and difficult questions and unfortunately it is rarely specified as a part of the evaluative process in most descriptions of educational evaluation.

Setting these evaluative standards is mostly a matter of thoughtful experience on the part of a teacher. This, of course, places the new teacher at a very serious disadvantage. In this case, a new teacher should use the experience of successful teachers and should consult pupils' texts, teacher guides, courses of studies, and any other material that may provide information about appropriate standards. It is important that the beginning teacher at least make an attempt to set forth these standards, no matter how crude they may be. These first attempts will start the teacher on the right path toward effective and valid

evaluation and will provide a basis for future revision. But at no time should these standards be considered fixed or absolute.

3. *Prepare the measuring devices.* After the evaluative standards have been set forth in the form of varying levels of performance, the third step in the process consists of constructing or finding a procedure that provides a measure of performance relative to each dimension of achievement cited in the standard. The various types of measuring procedures and specifications for their preparation have been thoroughly discussed in Chapters 3, 4, 5, and 6. In the case of Miss *X*, this step of preparing the procedure is easily recognized as the one where she formulates her questions and assembles them into a test. In her case, though, this was the first step, and we should see now that the test or other measuring device ordinarily should be devised in accordance with an already established evaluative standard. This means that, as the teacher constructs her test or measuring device, she should have beside her a set of evaluative standards in the form of levels of performance, and her test should be so designed that each level is adequately sampled. Such a measuring device will then permit the teacher to observe the performance of each pupil at each of the levels contained in her evaluative standards.

4. *Measure and evaluate.* Once the measuring instrument has been developed or selected, the fourth and final step in the evaluation process is simple. The instrument should be administered to the group of students involved, scores related to the standard, and qualitative symbols assigned. Each of the levels of performance in the evaluative standards should be represented by a symbol or by a single descriptive word. Ordinarily, when symbols are used, they are ordered in such a way that the first symbol represents the top level or gradation, the second symbol the next highest level, and so on down the line until the lowest level of performance is reached. Each student is assigned the evaluative symbol that represents the highest grade of performance achieved by the student.

In Miss *X*'s case, it is apparent that her students have been evaluated according to scores received on her test. Now, since she has no well-defined set of evaluative standards, she has to base her qualitative ratings upon some arbitrarily established distribution of ratings, or upon some arbitrary percentage of questions answered correctly. An example of the first alternative would be to decide that only 10 per cent of the group will be assigned the highest qualitative rating; therefore she selects the necessary number of students from those who stood highest on her test. An example of the second alternative would be to decide arbitrarily that those students answering less than 70 per cent of the questions correctly would be given an unsatisfactory rating. In the case of either alternative it is apparent that the ratings would have no clear-cut or consistent meaning in terms of performance.

TWO EXAMPLES OF EVALUATION

Now that we have discussed the four essential steps in the evaluation process, we look at two specific examples of an attempt to carry out these

steps. The first is an attempt made by a second-grade teacher. Her class has just spent a week studying a story and she is now ready to evaluate the accomplishment of her pupils.

In line with the first step in the evaluative process, the teacher has stated her objectives for this particular unit in terms of pupil behavior.

1. Ability to remember and understand the factual content of a story.
2. Ability to recognize and use the new words presented in the story.

The second step in the process asks that the teacher set forth her evaluative standards. This she does by describing the following levels of performance.

<i>Level</i>	<i>Performance</i>
I. Unsatisfactory	Cannot pronounce correctly the new words when seen written. Cannot answer questions about the most obvious facts contained in the story.
II. Fair	Can pronounce the new words correctly when seen written. Can answer questions about the most obvious facts contained in the story when the questions are phrased the same way as the statements in the story.
III. Good	Same as level II and in addition: Can underline the new words from among other words when the new words are spoken. Also has a rough idea of the meaning of the new words. Can answer more subtle questions about the content of the story when the questions are phrased differently from the sentences in the story. Also can recall the sequence of the story.
VI. Excellent	Same as levels II and III, and in addition: Can correctly use new words in sentences of his own construction drawn from his own experience, and can identify when the new words are improperly used. Can relate the content of the story to his own experience and can give plausible explanations of the events in the story.

Since the teacher is concerned with two objectives (or dimensions), she therefore needed to indicate varying levels of performance regarding the achievement for each objective. The way in which the performance levels are described above would indicate that the teacher has had already some experience with second-grade children and has done some serious thinking about what constitutes varying degrees of desirable behavior on the part of her chil-

dren. On the whole, the performance scale she used meets the criteria for evaluative standards outlined earlier in this chapter.

With the evaluative standards before her as outlined above, the teacher is now ready to take the third step in the evaluative process. This consists of developing a measuring device which will measure pupil performance at each of the four levels and relative to both of the dimensions in the evaluative standards. We shall provide here only a few examples of questions eliciting performance at each of the levels. At the second-grade level, the teacher will have to read the questions orally and the children will either answer orally or on a mimeographed sheet provided for them.

For Levels I and II:

(Teacher) "Read each of these words out loud."

1. hop NOTE: These words appear on the children's sheet and individually
2. shoe they are to pronounce each word as they see it.
3. walked
4. fun

Oral questions and answers:

1. What did Bill and Susan do when they played?
2. Bill said we must play in what kind of shoes?

For Level III:

(Teacher) "Draw a line under the right word in each box as I say it." 1. hop;
2. shoe; 3. walked; 4. our.

1. have <u>hop</u> how hot	2. show store <u>shoe</u> saw	3. <u>walked</u> wanted went walk	4. one out own <u>our</u>
-------------------------------------	--	--	------------------------------------

(Teacher) "Draw a line under the right words as I read each sentence."

1. Bill and Susan can ^{stop}hop.
2. Bill had a ^{hole}nail in his shoe.
3. The children ^{ran}walked in their new shoes.
4. Bill said, "We must play in ^{our}one old shoes."

Oral questions and answers:

1. What kind of store did Bill and Susan and Mother go to?
2. What did Bill do after he found the hole in his shoe?
3. Tell what happened after they left the store.
4. Can you show how Susan hops?

For Level IV:

(Teacher) "Make up a sentence about what you did or saw, using these words."

1. hole
2. walked
3. fun
4. looked

"Draw a picture about these words."

Oral questions and answers:

1. Why did Bill say, "We must play in our old shoes"?
2. What caused the hole in Bill's shoe?
3. Why didn't Bill's Mother scold him for the hole in his shoe?
4. Can you tell a story about when you thought you were going to be scolded and you weren't?

These are a few of the questions that the teacher in our example developed to measure performance at each of the levels set forth in the evaluative standards. Those designed for Levels I and II are keyed to minimal performance and are used to distinguish between unsatisfactory achievement and that which is barely satisfactory. Level III questions represent the next higher level of performance, and finally the last group of questions clearly is keyed to the top level of performance. It should be apparent that the construction of a measuring device becomes much simpler when the levels of performance in the evaluative standards have been carefully delineated. The teacher, of course, should construct a sufficient number of questions so that each dimension is well sampled at each level. (See pages 102-104 for a detailed discussion of sampling in test construction.)

The final step in the evaluative process consists of administering the test and assigning the qualitative symbol that represents each pupil's highest level of performance. Please notice that these symbols now have very specific meanings in terms of performance.

Further to demonstrate effective evaluation, we wish now to present an example from the upper elementary grades. In this one, the teacher wishes to evaluate his pupils' understanding of the following written passage.

The Great Law¹

Not long after his arrival in the New World, William Penn called a meeting at Upland of all the Swedish, Dutch, and English settlers. After making a short friendly speech, the Quaker leader told the people about the constitution, or set of laws, under which they were to be governed.

Penn's constitution was known as The Great Law. It provided that the colonists should be free to worship in any way they saw fit; that all settlers who paid taxes

¹ G. V. D. and J. V. D. Southworth, *Early Days in the New World*. Syracuse, N. Y.: Iroquois Pub. Co., 1950 (pp. 242-243).

had the right to vote; that every male member of any Christian Church might hold office; that all children should begin training for some trade or useful occupation at the age of twelve; that only two crimes, murder and treason, should be punishable by death; and that all prisoners should be put to work at some useful trade so that they might become good citizens. These last two provisions of Penn's Great Law are especially interesting in view of the fact that in England at this time there were over two hundred crimes punishable by death, and that no one had ever thought before of making a prison anything but a place in which to lock up people who had done wrong.

As a basis for judging the pupils' understanding of "The Great Law," the teacher prepares the following performance hierarchy:

<i>Level</i>	<i>Performance</i>
I. Unsatisfactory	Cannot answer questions about the most obvious facts contained in the story.
II. Fair	Can answer questions about obvious facts in the reading material.
III. Good	Same as level II and in addition: Can make comparisons and discriminations, and can formulate in own words definitions and illustrations of concepts presented in the reading material.
IV. Excellent	Same as levels II and III and in addition: Can give explanations and interpretations and can justify and predict on the basis of what is contained in the reading material.

With this variation scheme as her standard, the teacher may then develop an instrument for measuring performance at each of these levels. Some possible questions that refer to "The Great Law" are presented as follows:

For Levels I and II (Simple recall):

1. Who attended the meeting at Upland?
2. What crimes were punishable by death under the Great Law?
3. At what age were children required to start their occupational training?

For Level III (Illustrate, define, compare, discriminate):

1. How would you define treason?
2. Give an example of a person who would not be able to vote under this Great Law.
3. What is meant by the phrase "hold office"? Give an example.
4. What does "a trade or useful occupation" mean? Give an illustration of something that is and something that is not a trade or useful occupation.

For Level IV (Explain, justify, predict, interpret):

1. Why do you suppose it is required that a person be a taxpayer before he has the right to vote?
2. Suppose you were able to visit a prison in England and a prison in William Penn's colony at that time. What essential difference would you expect to find between these prisons?
3. Mr. Williams, a colonist, attempted to run for office but was disqualified. For what reason do you suppose he was disqualified?
4. Would you say that reducing the number of crimes punishable by death is an improvement? Give your reason.

These are some of the many questions that may be used to elicit responses indicative of each of the levels in the evaluative standard. The questions are written in short answer form but they can be changed to true-false, multiple-choice, or to some other guided response form if desired. Ordinarily a much greater amount of reading material is covered by a classroom test. This short passage was selected, however, to illustrate on a small scale how the evaluative process can be carried out effectively.

We can see in the two examples the importance of carefully prepared evaluative standards. They serve as guides for developing an effective measuring instrument. Furthermore, they give specific meanings to the qualitative ratings assigned. When a student receives a mark of "good" or a *B*, both the student and the teacher have a clear understanding of what the symbol means in terms of his performance.

Levels of Performance as Standards in Evaluating Achievement

Whether the standards are arbitrary distributions, teacher feelings, or given percentages of questions answered, levels of performance are, of course, involved in some way. Differences in rank in a distribution of test scores necessarily are somewhat a function of differences in performance. The feelings of teachers about pupils certainly are based on the performance of pupils and, in a sense, to answer a greater percentage of test questions is to perform at a different level. In this text, though, we are proposing that the levels of performance should always be explicit and that evaluative standards for achievement usually should consist of performance scales or hierarchies.

To help the beginner, we shall present a generalized scheme of variation in performance which is thought to be applicable to the *understanding* of any verbal subject. Since most subjects studied in school are verbal or have verbal aspects, the scheme should have widespread significance.

The use of defined levels of understanding as evaluative standards has not been extensive. Therefore this particular variation scheme should be considered as a beginning and not as a finished product. The levels are meant to be suggestive only. Depending upon the particular subject, grade, etc., the levels may overlap, they may omit important items, and some of them may be inappropriate.

*Performances Indicating Different Levels of Understanding
of a Given Subject*

<i>Level</i>	<i>Performance</i>
I	Imitating, duplicating, repeating. This is the level of initial contact. Student can repeat or duplicate what has just been said, done, or read. Indicates that student is at least conscious or aware of contact with a particular concept or process.
II	Level I, plus recognizing, identifying, remembering, recalling, classifying. To perform on this level, the student must be able to recognize or identify the concept or process when encountered later, or to remember or recall the essential features of the concept or process.
III	Levels I and II, plus comparing, relating, discriminating, reformulating, illustrating. Here the student can compare and relate this concept or process with other concepts or processes and make discriminations. He can formulate in his own words a definition, and he can illustrate or give examples.
IV	Levels I, II, and III, plus explaining, justifying, predicting, estimating, interpreting, making critical judgments, drawing inferences. On the basis of his understanding of a concept or process, he can make explanations, give reasons, make predictions, interpret, estimate, or make critical judgments. This performance represents a high level of understanding.
V	Levels I, II, III, and IV, plus creating, discovering, reorganizing, formulating new hypotheses, new questions and problems. This is the level of original and productive thinking. The student's understanding has developed to such a point that he can make discoveries that are new to him and can restructure and reorganize his knowledge on the basis of his new discoveries and new insights.

In using such a performance scale there is the assumption that the teacher knows pretty well what the students are reading and otherwise experiencing. If a teacher is not too well acquainted with what a student has read or done, then she may be misled into thinking that a student is performing at Level IV in being able to explain or predict while actually the student is operating at Level II, simply remembering what he had previously read or done. Furthermore, the student may be performing at Level V with some original thinking whereas the teacher might believe that the student is simply imitating or duplicating at Level I.

The fifth level of performance, original and productive activity, is the

ultimate goal of all education. At this stage the student is capable of independent work. Since most creative activity is spontaneous, the teacher cannot so easily solicit response at this level as she can at the other levels. It is also interesting to note that a teacher can ordinarily expect to lead a student through the first four levels of understanding but for the fifth level the student must be his own guide.

As a final remark on the evaluative process, we must say that much more study is needed in developing general variation schemes for quality of performance that can serve as a basis for teachers to develop their own evaluative standards for their own particular situations. Substantial progress has been made toward setting up carefully defined objectives for various school activities. However, there has been too little done toward establishing differential levels of performance in the attainment of these objectives.²

The Function of School Marks

At first glance, it seems that if the evaluation process has been carried out properly, there should not be much of a problem in reporting the results. On the contrary, we shall find in this section that the reporting of school marks is a very vexing problem even with effective evaluation. The latter, though, is prerequisite to any possibility of success in reporting pupil progress, as Wrinkle has testified:

Finally, almost ten years later, we discovered that we couldn't report intelligently unless we first evaluated intelligently, and that we couldn't evaluate intelligently unless we knew what we were trying to do (15:3).

Before we get involved in the details of reporting practice, let us consider the several functions that marks are thought to serve. We might do this by asking who are interested in marks and why.

Certainly the teachers are interested in them, although their job might present less frustration if report cards were discarded. For the most part, they use them in valid ways—to tell their pupils what they think of their work and to determine what has been their achievement in other subjects and in earlier grades. In addition some teachers like to “motivate” pupils by promising *A*'s for good work and *F*'s for slothfulness. The few who are vindictive can punish pupils who have bothered them by giving such pupils low grades. If they are guidance-minded, they may attempt to use marks to promote pupil adjustment.

The pupils, themselves, are of course interested in the grades they receive. They wish to know how well they are progressing and in what areas and activ-

² The *Taxonomy of Educational Objectives* (2), cited in Chapter 2 for its analysis of dimensions of knowledge, includes as well an attempt to establish levels of performance relative to these dimensions. This book is highly recommended to all teachers who wish to set valid evaluative standards in their teaching. It might be well to compare the hierarchy of performance levels proposed by the college examiners with that proposed by the authors here.

ities they need further development. They need the information provided by marks so that they can make realistic educational and vocational plans for the future. And, of course, in secondary schools there are honor rolls, athletic eligibility, and many other things that hinge on school marks.

The parents of the pupils are extremely concerned about the marks given their children. Among the reasons for their concern are genuine interest in their children's progress, the prestige value of high marks, the "loss of face" attendant upon *F*'s, and their need, like that of their children, to have some basis for educational and vocational planning.

All superintendents and principals must be interested in marks for administrative reasons. They form an important part of the cumulative record for each pupil. Grades are used as a basis for promotion or nonpromotion, and for the determination of honor rolls. High schools are interested in seeing if a student's elementary marks justify his enrollment in given subjects. Colleges often use high school marks as a basis or a partial basis for admission. Schools also must be ready to provide information on the academic standing of a student to a potential employer when requested. Consequently, in view of all these administrative uses of grades, marks are well-nigh indispensable to principals and superintendents.

A final group who may wish to know a pupil's school grades is composed of prospective employers. Over-all high marks suggest a capable worker and over-all low ones, a poor worker. Moreover, certain types of employers are interested in grades in given subjects: the insurance firm in typing and book-keeping, the sales firm in public speaking, the chemical plant in science and mathematics.

So far we have identified five groups of persons who are concerned about grades, namely: the teachers, the pupils, the parents, the school administrators, and the potential employers. The reasons for their interest are in effect the functions of school marks. In summary, these seem to be:

1. Indicate academic standing and competence.
2. Facilitate instruction and guidance.
3. Provide motivation for learning.
4. Serve as a basis for future planning.
5. Serve administratively for placement, promotion, certification, admission, and for permanent records.
6. Serve as predictors of school and vocational success.

From this brief résumé of who are interested in grades and for what reason, we can readily see why marking has become so complicated and confused. There are many different persons and many different functions to be served by the same marks. Yet the type of mark that best serves one group or function may least serve another.

Marking and Reporting Practices ³

In day-by-day and week-by-week contact between pupils and teacher, the evaluation process outlined earlier in this chapter should serve effectively. The specific objectives for the classroom activities would be carefully defined, an evaluative standard consisting of levels of performance for each objective would be prepared, a device for measuring performance at each level would be developed and administered, and evaluative ratings would be assigned accordingly. For each day and each week, the pupils would know what progress has been made and what has been accomplished. The pupils could readily tell at what level they were performing and what needs to be done in order to perform at a higher level. The teacher could know how effective the instructional activity has been and what areas need further emphasis. The qualitative ratings assigned by the teacher would have clear-cut meaning and there would be a continual two-way communication between teacher and pupils. In this case the functions of evaluation would be effectively served.

Trouble comes when the results of classroom evaluation are to be reported to parents, school authorities, and to the public, particularly in the form of a single mark at the end of a course. These persons, of course, have not seen day-by-day and week-by-week evaluations made in the classroom. They tend to be unaware of the complex nature of achievement in any subject, of the tentative nature of each teacher judgment, and of both the teacher's objectives and his standards. All they see is a semester grade and whatever meaning they give it has an indeterminate error.

SINGLE LETTER MARKING

The practice of using a single letter, usually *A, B, C, D, or F*, as a qualitative rating for a student's performance in a subject or activity has been seriously questioned by many. Their criticisms seem to center around the following points:

1. Each subject and each activity in school consists of many diverse aspects and each requires a variety of different skills and abilities. It hardly seems reasonable that a single mark could truly represent quality of performance in all these aspects. At best it must represent an average of qualitative ratings for all of them.

2. A single mark actually communicates very little, allows for no explanation, and provides only a limited basis for action. These marks scarcely indicate the pupil's strong points and weak points, and certainly they give hardly any indication of the pupil's potentiality. Without this information there is little on which to base any positive action.

3. Letter marks are often construed as rewards and punishments for the pupils and as prestige symbols by their parents and the public. When thus

³ Samples of several report cards are shown in Appendix C.

construed, grades have become ends in themselves, something to be achieved for their own sake instead of serving to facilitate learning. We should not have to look far to find children who are going through "motions" in the classroom just to achieve high marks and who are not concerned about learning anything. This is extrinsic motivation at its worst.

4. There is evidence that grades may have inconsistent meanings. Studies have shown that different teachers will rate differently identical performances and that the same teacher will rate the same performance differently on different occasions. Since no common meaning has been established for these marks, this certainly makes it difficult for parents and other outsiders to determine just what the marks do mean. The prevalent unreliability of marks also results in their limited use for predicting future academic success, for placement in courses, and in general as a basis for future planning.

Attempts have been made to de-emphasize the role of marks and thus to remove the possibility of their becoming ends in themselves. This has generally been done by reducing the number of letters used in marking and by changing their meaning. For example, the *A, B, C, D, E, F* letters are replaced by the letters *S* and *U*, *S* for satisfactory and *U* for unsatisfactory. This has certainly de-emphasized the role of marks and has relieved the pressure of competition on the pupils. On the other hand, the problem of communicating the accomplishments of the pupils has only been aggravated. The use of only two letters provides less information regarding the accomplishment and progress of pupils. In general, schools that have tried the experiment of reducing the number of letters used as marks have reported that the results were unsatisfactory, particularly in regard to their responsibility for providing information to parents and others about the competence of their pupils.

ANALYTIC EVALUATION

Schools have also sought answers to the objection that single letter marks actually convey little information and provide little basis for action or proper interpretation. One approach to this problem has been to list specific dimensions or aspects of a subject or activity and then to provide qualitative ratings for each. For example, in social studies achievement the following dimensions could be listed:

1. Developing a special vocabulary.
2. Recognizing events and their chronological relationships.
3. Reading and interpreting graphs, tables, and maps.
4. Locating and evaluating sources of information.
5. Analyzing social problems.

This procedure has much to commend it as a way of providing more information, but some drawbacks exist. For one thing, the number of specific components for each subject or activity can become very large and hence the amount of measurement involved becomes overly extensive. Then the job of

evaluating each pupil becomes too time-consuming and complex for the teacher. Another but smaller difficulty lies in wording each aspect so that parents and others can understand what is being rated. If these difficulties can be overcome, the use of a check list of components for each subject has great potentialities.

PARENT-TEACHER CONFERENCES

Another procedure used by schools in order to overcome the limited communication of a report card is to establish periodic parent-teacher conferences. This procedure has the advantage of a two-way personal interview in which the status of the pupils can be discussed informally and completely. A series of conferences of this sort requires careful scheduling and they are very expensive of time. Moreover, such conferences require a minimum level of competence on the part of the teacher if they are to be successful. Experience with the procedure has indicated that, in general, parents are initially enthusiastic about the conference but as soon as the newness wears off their interest decreases and subsequently their attendance drops off. Some schools use an informal letter to the parent in order to offset some of the difficulties encountered with the parent-teacher conference; however, this again may make evaluation an excessive burden to the teacher.

ROLE OF STANDARDIZED TESTS IN EVALUATION

Standardized tests have been particularly useful in helping to relieve the heavy import carried by school marks. These tests have been especially useful for predicting future academic success and for placement in courses. Colleges are more and more relying on a combination of standardized test scores and high school marks as a basis for determining those who are most likely to succeed in college, rather than relying solely on high school marks. Standardized tests are also useful in helping teachers check their grading and thereby give some consistency to their marks. There is an extreme point of view, with which the authors disagree, that such tests should be the sole determinants of final marks in courses. One can readily see how these tests may supplement the regular evaluation process but it is difficult to see how they can supplant it.

WHAT THE LETTER MARK MAY REPRESENT

The practice of giving a single mark as a qualitative rating for a course seems to be deeply rooted in our present-day educational system and there seems to be no indication that this practice will be drastically changed in the near future. This raises the question, then, of what the single mark should represent. In assigning a single mark to a pupil, should the teacher give consideration to the pupil's effort, capabilities, initiative, co-operation, and to the effect it might have on the pupil and his parents? In a recent survey of 53 California school systems of various sizes, 27 school systems reported that

they believed that the basis for awarding subject-area marks should be a composite of subject matter achievement, co-operation, effort, and initiative (4).

This is contrary to the principles set forth earlier in this text, that any symbol is meaningful only when it refers to one dimension. This means that the single course mark should be devoted only to evaluating status as to competence in the subject. All other considerations should be rigorously excluded. This, of course, is much easier said than done, for we teachers *are* human and the temptation is strong to consider the effect the marks will have on the pupils and to use marks for rewards and punishment. In order to reduce this temptation and also to provide for the evaluation of "citizenship," some schools include certain nonacademic dimensions on their report cards and rate them separately. These include such items as citizenship, work habits, attendance, co-operation, conduct, effort, responsibility, and attitude.

Criteria for a Marking and Reporting System

Determination of a thoroughly satisfactory report card and/or marking system thus seems to lie in the future. However, in our criticism of current practices there have been implied certain criteria for more valid marking and reporting. Some of these may be met by carefully handling any conventional system. To accommodate to others will require procedural innovations, released teacher time, additional training on the part of teachers, etc.

1. *Marks in given subjects should be based on extensive and comprehensive measurement.* A pupil's achievement in a subject—arithmetic, reading, biology, or shop—is his performance over a span of time and consists of many dimensions. To measure it only once (say, a final examination), to measure only one dimension (say, memory of facts), or to use but one measuring procedure (say, a guided response test) is to assume that *any* pupil performance is clear evidence of all the pupil's performances. This, of course, is known to be a false assumption so, to be valid, marking must be based on extensive and varied measurement of all essential aspects of the achievement in question.

2. *The marks must symbolize a comparison between pupil status and known and fair standards.* As we have emphasized throughout this chapter, the *sine qua non* of effective evaluation is the evaluative standard. Without one, evaluation must be subjective, capricious, and erratic. For the teacher to have a standard, though, is not enough for pupils and parents. They wish to know what it is and to know that it is fair. Discussion with pupils may serve to inform them but parents should have the standard(s) in writing. This may be on the face of the report card or on a separate form, but it should accompany transmission of grades. Fairness in a standard hardly needs justification but in gaining it two considerations are critical. As we have asserted, the maturation of pupils in the grade in question must be a limiting factor in setting standards. Moreover, standards should be derived from the performance of pupils rather than from the performance of the teacher. The latter's

performance almost by definition is sure to outstrip that of the best pupils, and consequently, if it is used as a rigid standard, even the best pupils may receive unduly low marks.

3. *The marking and reporting system should be somewhat diagnostic.* Since learning is best promoted by detailed knowledge of right and wrong actions, generalized evaluations do little to promote it. The basic components or dimensions of performance in English, science, history, etc., must be marked separately if the pupil is to know what to exploit, what to forget, and what to correct. Earlier, on pages 29-32, 93, we cited several examples of subject breakdown. In the next section of the book, the necessary analysis is described for most school subjects.

4. *Marks and reports must help the pupil assess his accomplishments realistically.* Much has been said about assuring success for pupils and minimizing frustration for them. With this stand we agree, but we think that great care should be taken in extending the principle to marking practice. For a student to receive a mark that symbolizes *more* accomplishment than he has actually achieved is, we feel, a disservice to the child. He will tend to overrate his competence and thus may be sharply disappointed when he tries it out in a less benevolent circumstance. On the other hand, good pupils who are marked low simply to prod them to more accomplishment may for that reason underrate their talent and fail to aspire as high as they should.

We consider that the primary function of marks is to communicate a realistic evaluation. In doing this it may be appropriate to mark both on the basis of achievement and on gain or learning. It seems that for most subjects pupils profit from knowing how much they can do at the moment *and* how much progress they have made, so a separate mark might well be given for each type of evaluation. Certainly, the pupil and parent should not be left to guess which is the basis nor should some pupils in the same class be marked on gain and others on level of achievement.

The frequent practice of differential marking according to ability, if not carefully administered, can produce unrealistic self-evaluations on the part of pupils. In this procedure each pupil is judged according to how much may be expected of him. If the dull pupil lives up to expectation, he receives the same mark as the bright pupil who lives up to expectation. Yet, by definition, the pupils' achievements are entirely different. Some of the misleading aspect of "marking on ability" can be avoided by adding subscripts to the grades, i.e., *x* for bright pupils, *y* for average, and *z* for dull ones, or some such.

The primary virtue of ability marking is its avoidance of competition among ill-matched pupils. We feel that the use of dual marks, one for absolute achievement and one for gain, will serve to minimize this unfair competitive aspect of marking and still be realistic. The poorest pupil feels "success" over his progress but still recognizes the low level of his competence. The brightest pupil, on the other hand, can be reprimanded by an *F* for his lack of progress but still be told by an *A* that his achievement is nonetheless of high order.

5. *The symbols and the report form must be understandable to pupils and parents.* Obviously the meaning of marks should be clear and constant from year to year and from teacher to teacher. Such technical terms as percentile and standard score should be avoided. All entries on the forms used must be immediately meaningful or there defined.

6. Finally, *the preparation and administration of report cards or the alternatives of letters and conferences should entail no more time than is available for the purpose.* Teachers should not be expected to work overtime to achieve effective reporting. Even if they are willing to do so, their efficiency will drop and much of the value of the procedure will be vitiated. If school boards and administrators cannot provide time enough in the regular school day for teachers to adhere to some of these criteria that involve additional time, it should be anticipated that marking and reporting simply will fall short of the criteria.

Summary

Evaluation is the process whereby the quality or value of anything is determined. Usually, this involves a comparison of the status of the thing in question with a standard appropriate to a determination of its value. Among the standards in use in schools are purposes, averages, arbitrary conceptions, and customs. A performance scale is considered the best type of standard for evaluating achievement.

To be satisfactory, evaluative standards should:

1. Present a clear-cut variation of quality or value with different categories of value well defined;
2. Be practical for use and expressed in terms appropriate and favorable to the best procedures of measurement;
3. Be consistent with scientific findings about learning and child development; and
4. Be consistent with the values of our culture.

Evaluations are expressed in much the same forms as measurement, namely, by classificatory and descriptive symbols, by indexes of rank, and by scale numbers. The last type of evaluative symbol is rarely possible in education. Whatever the form of expression, it is essential that the symbol be related to unambiguous and appropriate statements about the quality gradations they represent.

To accomplish valid evaluation of pupil achievement it is necessary first to determine what dimensions of achievement are to be judged. Second, a standard should be selected or prepared appropriate to the dimensions. In the third step, needed measuring devices should be procured or prepared and designed according to the levels contained in the evaluative standard. Finally, measurement must be performed and the evaluations made and communicated to pupils and others who need to know them.

A generalized standard proposed for use in evaluating pupil understanding of any subject is composed of five levels of performance: I. imitating, duplicating, repeating; II. Level I plus recognizing, identifying, remembering, recalling, classifying; III. Levels I and II plus comparing, relating, discriminating, reformulating, illustrating; IV. Levels I-III plus explaining, justifying, predicting, estimating, interpreting, drawing inferences; and V. Levels I-IV plus creating, discovering, reorganizing, formulating new questions, hypotheses, and problems.

School marks, on tests, papers, homework, and the semester's work, are symbols of teachers' evaluations of pupil achievement. As such, they serve to facilitate instruction and guidance, motivate study, serve as a basis for future planning, for placement, promotion, and admission, and for prognosis of school and vocational success. Traditional practices in marking have many weaknesses but no entirely satisfactory new method has yet been devised. Among the conditions essential for a valid marking and reporting system are these items: Marks should be based on extensive and comprehensive measurement. They must symbolize a comparison between pupil status and known and fair standards. The system should be somewhat diagnostic. It should help the pupil assess his accomplishments realistically. The meaning of symbols must be clear to pupils and parents. Finally, the marking and reporting process should put no excessive burden on the teacher.

EXERCISES

1. Compare the hierarchy of performance levels as presented by Bloom (see bibliography) with the hierarchy as suggested in this text.
2. Prepare a generalized set of evaluative standards in the form of successive levels of performance that would be generally applicable to your teaching area.
3. Select a specific objective in your teaching area, develop a set of evaluative standards for this objective, and construct a few illustrative items that would be contained in your measuring instrument.
4. Examine carefully at least three report cards that are in use. (You may refer to the report cards presented in Appendix C if you are unable to obtain your own copies.) Make a critical analysis of these report cards in the light of the discussion of this text.
5. Develop what you believe to be an ideal report card for a course you teach or plan to teach.
6. Suppose you have been selected to serve on a committee that plans to study the marking practices in your school. What recommendations would you make to insure uniform marking throughout your school?

BIBLIOGRAPHY

1. Bloom, B. S., and Heyns, I. De V., "Development and Application of Tests of Educational Achievement," *Review of Educational Research*, 26:71-88, February, 1956.
2. Bloom, Benjamin S. (ed.), *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*. New York: Longmans, Green, and Co., 1956.
3. Elsbree, Willard S., *Pupil Progress in an Elementary School*. New York: Bureau of Publications, Teachers College, Columbia University, 1943.
4. Higman, William C., "An Evaluation of Pupil-Progress Reporting in Selected California High Schools," unpublished Master's Thesis, Sacramento State College, 1952.
5. National Society for the Study of Education, *The Measurement of Understanding*, 45th Yearbook, Part I, University of Chicago Press, 1946.
6. Odell, C. W., "Marks and Marking Systems," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 711-717.
7. Reitz, Wilhelm, "Higher Mental Processes," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 540-551.
8. Roelfs, R. M., "Trends in Junior High-School Progress Reporting" *Journal of Educational Research*, 49:241-250, December, 1955.
9. Russell, I. L., and Thalman, W. A., "Personality: Does it Influence Teachers' Marks," *Journal of Educational Research*, 48:561-564, April, 1955.
10. Scates, D. E., "Some Problems Connected with Evaluation," *Journal of Educational Research*, 45:599-608, April, 1952.
11. Smith, E. R., and Tyler, R. W., *Appraising and Recording Student Progress*. New York: Harper Bros., 1942 (Vol. III of *Adventure in American Education*).
12. Strang, Ruth, *Reporting to Parents*. New York: Bureau of Publications, Teachers College, Columbia University, 1947.
13. Thorndike, R. L., and Hagen, Elizabeth, *Measurement and Evaluation in Psychology and Education*. New York: John Wiley and Sons, 1955. Chap. 17.
14. Werkmeister, W. H., *The Basis and Structure of Knowledge*. New York: Harper Bros., 1948.
15. Wrinkle, William L., *Improving Marking and Reporting Practices in Elementary and Secondary Schools*. New York: Rinehart & Co., 1947.

SECTION II

CUSTOMARY USES OF MEASUREMENT AND EVALUATION IN EDUCATION

OVERVIEW

The general processes of behavioral measurement and some problems involved therein have been presented in the first section of this text. In addition some attention has been given to the evaluation of pupil achievement and to effective modes of reporting evaluations to pupils and parents. Our purpose has been to develop a generalized scheme for measurement and for evaluation that is valid, and will make for more efficient evaluative practices in schools. Now we wish to apply this scheme to the tasks of educational measurement and evaluation that teachers and administrators must perform.

Never before in history have schoolmen been so interested in accurate assessments of achievement and the factors related to it—intelligence, personality adjustment, reading. In the complex origins of this circumstance, two factors are paramount. In the first place, educational practice has at last been given some scientific underpinning. Out of the laboratories of Wundt, Watson, Thorndike, Skinner, Yerkes, and others have come experimental findings about learning that support a demonstrably effective methodology of instruction. As part of this scientific movement in education, both the factors that affect learning in the schools and the outcomes of instruction have been expected to yield up their secrets to the probing of measurement. Moreover, the analogy of the laboratory and the classroom, of research and of everyday instruction, has largely been accepted. If precise measurement has been applied to motivation, practice, error, learning, and retention by the researcher, should it not also be applied by the teacher?

As we know, of course, the great promise of objective measurement in education has not been entirely fulfilled. The tests have not been completely reliable nor has their focus been exact. Teachers and administrators have lacked the time and skill to do all the things the psychometrists have said must be done. And, unfortunately, many of the testing practices of the 1910's and

1920's have been found to rest on erroneous statistical and psychological assumptions. However, to offset these misadventures, there has been continued refinement of testing procedures and instruments. With decreased dependence on the "objective" test (guided response) has come increased efficiency in the use of essay examinations (free response). And a great variety of novel measuring techniques has been developed in pace with the school's widening interest in understanding, critical thinking, socially desirable attitudes, and the like.

Of equal force, perhaps, in increasing the importance of measurement in the schools are two relatively new aspects of American public schools, guidance and ability grouping. The counselor requires IQ's, reading grade placements, percentile ranks on interest tests, and cumulative records if he is to guide his students toward proper vocational goals and to help them with their problems of personal adjustment. For efficient ability grouping, whether within a class or among classes, it is necessary to have reliable information about the general ability and the subject readiness of each pupil.

If the measurement and evaluation schoolmen now must perform is to be efficient, it is thought that it should be consonant with some rational and systematic approach. Testing and marking is now too complex, too ramified, and too important in the lives of pupils for it to be practiced casually. The scheme presented in the first section is designed to produce more efficiency and validity in the evaluation of pupils in various subjects. It has been developed in full knowledge of the many problems involved in marking assignments, scoring tests, observing laboratory performance, and giving grades. It offers no pat solutions or panaceas for these problems, but it does offer a rational approach to their solution.

In this section attention will be given in order to the Language Arts and its many specific subjects, to Social Studies, to Science and Mathematics, and then to the performance-activity subjects of Music, Art, Physical Education, etc. For each of these subjects, measurable dimensions are discussed, appropriate measuring forms and procedures are described and illustrated, evaluative standards are presented, and marking practices are noted. The measurement of intelligence is examined subsequent to this, as is measurement applied to personality and character. The section and the book conclude with a discussion of school-wide testing programs.

No attempt has been made in the section to deal with every possible aspect of measurement and evaluation for every school subject. To some extent, our coverage of subjects is a function of the amount of research that has been devoted to their measurement. In addition, we have tried to consider the usual training programs for different types of teachers. Thus we have dealt less exhaustively with the highly specialized subjects (art, music, physical education, etc.) than with the more general ones, since the training of teachers for these special subjects usually involves specific instruction in measurement and evaluation as applied to the subjects. Then for each subject area we have tried to emphasize the procedures most significant for the area. Finally, we have not

attempted to repeat discussions of procedure and principle where their discussion in one context may be easily understood and applied to another.

In view of this, it may be advisable to study all of certain chapters and portions of others in addition to the one representing your specialization. To help you determine what you will want to read, we have outlined below the special emphasis in each of the chapters on subject areas and portions that may have general significance. The three remaining chapters—Intelligence, Personality and Character, and School-wide Testing Programs—are, of course, pertinent to teaching at any level or in any subject.

Special Emphases	Passages of General Significance
<i>Chapter 10—Language Arts</i>	
Guided response techniques.	Reading, pages 222–236.
Standardized tests.	Composition, pages 238–250.
<i>Chapter 11—Social Studies</i>	
Free response techniques.	Evaluative standards, pages 272–274.
Analysis of written products.	Citizenship and psychological grading, pages 275, 276.
	The prototype study of all phases of measurement and evaluation in an eighth-grade class, pages 278–291; particularly the over-all plan, pages 282, 283, and attention to study habits, pages 286–289.
<i>Chapter 12—Science and Mathematics</i>	
Guided response techniques.	Evaluative standards, pages 300, 313–319.
Performance-scale evaluative standards.	
<i>Chapter 13—Performance-Activity Areas</i>	
Observation techniques and analysis of graphic products and artifacts.	Performance tests, pages 339–354.
Performance tests and performance rating scales.	

CHAPTER 10

LANGUAGE ARTS

Language, in its various aspects, is the *modus operandi* for learning in the schools. Pupils and teachers talk and listen, books are read and papers written with a view to learning given things—arithmetic, history, governmental forms, social problems, science. In addition, the outcomes in these subjects, the things learned, are themselves language in many cases—the basic symbols or vocabulary of a subject, concepts, statements of fact and relationship, theories, etc. Moreover, the skills of language are objects of instruction in the elementary grades. In secondary grades, to these are added the artifacts of language, novels, poems, essays and short stories; plays, new stories, biographies, and propaganda.

In consequence of this fact that our schools are primarily language schools, in one form or another, the measurement and evaluation of language arts phenomena is a matter of great concern in all school grades. In the lower grades where the development of rudimentary language skills are of critical importance, primary attention is given to reading, handwriting, and vocabulary. In the upper grades the measurement of these phenomena per se is less significant and spelling, punctuation, grammar, and composition are the phenomena more often evaluated. Then, in secondary school and in college, specialized aspects of language become the major objectives of evaluation—public speaking, literature, creative writing, foreign languages, dramatics, and journalism.

Our plan of presentation in this chapter is first to describe some general features of measurement and evaluation in language arts and then to discuss their application to each language skill or subject. In both the general and particular treatments, the organization of the first section of the book will be followed. The dimensions of the phenomena will be presented, then the forms and procedures of measurement that are most relevant. After this, attention will be given to applicable evaluative standards and finally to problems of marking and reporting.

GENERAL FEATURES

Measurable Dimensions

Each of the language arts subjects has several measurable dimensions. Collectively, these have but two common characteristics: they are largely symbolic behaviors and, for the most part, they are overt behaviors. Being overt, the problems of inferred dimensions and constructs are rarely encountered (see pages 26–28). Because they are symbolic the matter of clear definition is perhaps the most critical condition to be met in preparing to measure them.

When an attempt is made to measure language arts achievement as a whole, the elements of language arts are themselves the basic dimensions for which test scores or ratings are sought. Reading, spelling, vocabulary, grammar, and knowledge of literature, several or all, are the focus of general language achievement tests. A pupil's status with respect to the particular dimensions of each element usually is not assessed separately.

Forms of Measurement

While some of the various language arts entail special measurement expressions, nearly all are limited to the precision of classification and rank symbols. In some cases, scores from teacher-devised tests may be converted to percentile ranks and nearly all standardized tests yield percentile ranks or some other index of rank in a large population.

Even though the differences among pupils in most of the dimensions of language arts are a matter of degree or of continuous variation, scale measurements are hardly applicable to them. Scale measures require fixed points of reference and regularized units of difference, and these are hard to contrive for language arts. It is true that so-called "scale" scores are produced by some tests and many rating "scales" for language arts are in existence. However, the use of the term usually is figurative rather than literal. A pupil's rate of reading may, of course, be expressed in words-per-minute, a scale index, as can his rate of writing or talking.

Procedures of Measurement

Appraisal of progress in language arts subjects, more often than not, is based on informal and even incidental observation and product analysis. The marks that teachers give pupils are heavily influenced by what they see and hear as pupils express their ideas in discussion, study, write papers, and use the library. The pupils' paragraphs and compositions, their fill-ins of work sheet pages, and their notebooks are inspected by teachers as a further basis for marking. Thus, the unreliability of subjective judgments is a problem in language arts evaluation. But, on the other hand, the inherent validity of direct observation and product analysis is a compensating boon.

Both free and guided response tests are used widely, particularly for aptitude, instructional classification, diagnosis of difficulty, and for other purposes that do not involve marks and report cards. In addition to his own tests, a teacher of a language arts subject has available a large number of published standardized tests. Table 9 shows the great number listed in the *Fourth Mental Measurements Yearbook* (1).

TABLE 9
Published Tests listed for Language Arts
Areas in the *Fourth Mental Measurements Yearbook*

English in general	30
Composition	2
Literature	18
Spelling	15
Vocabulary	6
Reading	36
Oral	1
Readiness	7
Special fields	5
Study skills	11
Total	131

The general advantages and disadvantages we noted for standardized tests (page 123) hold true in language arts. Their use permits comparisons with a larger population of pupils than locally devised tests and their construction is likely to be superior; but they may not suit either the objectives or content of a particular class and they often connote more "scientific" measurement than they attain. As we shall see in succeeding sections, standardized tests are far more valid for some language arts subjects than others. In reading, for example, they are well-nigh indispensable but in composition and literature their worth is often questionable.

Evaluative Standards, Marks, and Reporting

In one way, evaluation in language arts subjects may be fairly business-like and objective. Dictionaries and style books have recorded accepted word meanings, spellings, punctuation and, to some extent, usage. The teacher may judge papers submitted by pupils adversely as they differ from these standards and commend them as they approximate the standards. In another way, though, the evaluation of language arts phenomena may be vague, subjective, and even arbitrary. The other standards by which pupils' reading, writing, and talking are judged are such things as ability level expectancies, course of study objectives and standards, and teachers' opinions of what is "proper" for a fifth grader or a twelfth grader. These standards often are ill-defined; they vary from region to region and school to school; they may not be written

down; and, in many instances, they derive from custom or precedent rather than utility.

Marks rarely are given for language arts as a subject of study (it is a collective term) but rather for reading, spelling, English, journalism, etc. Practices of marking in language arts subjects display all the variety and common tendencies we noted for marking in general in the previous chapter.

As in other broad subject areas, many school districts have formulated evaluative standards, policies, and techniques for district-wide and multigrade use. These activities are considered extremely valuable, even though the evaluation plans achieved may have technical flaws. Only through such broad and cross-grade planning can any sort of comparable and accumulative evaluations be attained for the pupils who go through the school system. If evaluations are not comparable from grade to grade and are not cumulative, pupils are bewildered and teachers lack essential information.

READING¹

Reading Readiness

Even before children open the first preprimer in the first semester of the first grade, measurement usually has begun in reading. It has been known for a long time that a given level of maturity must be attained by a child before he is capable of learning to read. While important variation is to be noted in individual cases, by and large a child is ready to profit from instruction in reading only when he has the mental age of six years, six months, and possesses the social, emotional, physical, and experiential attributes that tend to accompany this degree of mental development in middle-class native-born American children. Consequently, perhaps the first measurement task that confronts a first-grade teacher² is to gauge her pupils' readiness for reading according to this general criterion.

DIMENSIONS OF READING READINESS

No given list of dimensions of reading readiness has yet been agreed to by authorities in reading. Mental maturity, visual and auditory acuity, competence in oral language, and experience relative to the objects and ideas in first readers seem to be the most commonly cited factors. The many published

¹ Measurement of reading status, aptitude, and weaknesses is intimately related to instruction in reading. The many current textbooks on reading methodology contain extensive material on measurement and should be consulted by any person who wishes to make a thorough study of measurement and evaluation in reading. Moreover, teachers' manuals that accompany a series of readers usually suggest appropriate evaluative devices and may even present ready-to-go tests.

² That this section is presented from the point of view of a first-grade teacher does not mean that readiness is of no concern in later grades. For any child who has yet to read readiness should be gauged at whatever grade.

tests of reading readiness differ among themselves as to the dimensions they attempt to measure. As a rule, the tests are directed toward certain factors that also are measured by intelligence tests but add others that relate particularly to reading. The dimensions that two rather different types of readiness test purport to measure are listed below.³

<i>Metropolitan</i>	<i>Murphy-Durrell</i>
Word meaning (oral)	Auditory discrimination (likenesses and differences of letters and words)
Sentence meaning (oral)	Visual discrimination (likenesses and differences of letters and words)
Information	
Visual discrimination of object likenesses and differences	
Knowledge of number	Rate of learning of vocabulary
Ability to copy	

FORMS AND PROCEDURES OF MEASUREMENT FOR READING READINESS

By definition, reading readiness is a composite of many variables, both immediate and historical; and status relative to each of them usually is expressed in a different form. For mental maturity it is customary to use Mental Age, a rank symbol. Visual and auditory acuity may be expressed in Snellen chart ratios and in watch tick or numbers heard distances. They can be, but usually are not for school purposes, expressed by the precise scale numbers of oculist's instruments and an audiometer test. Oral language ability usually is merely described in words classified or, at best, is reflected in the percentile scores of a readiness test. Experience is necessarily gauged by verbal description and by rough classifications.

Because of a lack of uniform indexes of measurement for the several variables of readiness, no single measure may be assigned to a pupil's readiness as a whole. Hence, the measurement of reading readiness is largely a diagnostic procedure. It is like a physical examination or a case study. It is not like determining a child's weight or even his intelligence or facility at spelling. While, as we shall see, reading readiness tests may yield single scores, they do not by any means appraise all the important dimensions of readiness.

The procedures by which this diagnosis is to be accomplished are as varied as the forms of measurement to be applied to its components.

Mental Maturity. The best measurement of mental maturity is, of course, a carefully administered intelligence test (see Chapter 14 and Appendix B). A group test should suffice to detect those who have the requisite Mental Age but individual tests should be ordered to verify findings for pupils with Mental Ages apparently too low for reading. If intelligence tests cannot be used for some reason, recourse is necessary to the first-grade teacher's observation of the pupil during the first weeks of school and possibly parent conferences and

³ These tests along with others are described in Appendix B.

neighborhood visits. While this latter means can yield no number score, it can yield a verbal estimate of the child's brightness relative to members of his peer group. With experienced teachers, estimates of relative intelligence can have surprising accuracy.

Sensory Acuity. The time-honored test of visual acuity is the Snellen or similar vision chart. However, many children will have 20/20 vision so far as the chart is concerned, but still are not able to focus properly on reading material 18 inches from their eyes. The chart test measures visual acuity at a distance while reading requires acuity of near vision. Children, as we know, do not develop the ocular adaptation for fixating clearly on small objects close at hand until about the time they enter school. Thus, the ordinary Snellen chart eye examination should be supplemented by an oculist's examination for any child with normal distance vision who has any difficulty with seat work. It presently is customary and advisable to refer to an oculist any child whose distance vision is appreciably less than normal.

Again, for hearing, there is a traditional test, the watch tick or whispered numbers. But as with vision, the gross results of the traditional procedure often fail to show the differential aspects of hearing that are important for speech. Consequently, use of a variable frequency audiometer is desirable. This instrument broadcasts tones whose pitch, as well as amplitude, may be controlled by the examiner. With it a child's ability to hear may be determined for all audible frequencies. School nurses are being trained today in the use of the audiometer, as are school psychometrists and speech clinicians.⁴

Oral Language. The third general dimension of concern in reading readiness, oral language ability, may be appraised in several ways. Many reading readiness tests contain vocabulary and sentence-meaning sections. Casual listening to pupils' conversation at recess and during activity periods usually will serve to detect those whose language development is accelerated and, as well, those who may be retarded in speech. For the latter, more systematic and detailed observation is in order. Storytelling sessions are excellent devices for revealing pupils' oral language facility and they serve other instructional purposes as well. A more contrived but fairly specific procedure is to ask pupils to tell what certain words mean. The words can be selected from primers or from word lists and thus a child's oral familiarity with the words he must learn to read can be determined rather precisely.

Experience. A child's general experience with adults, peers, and books has a large bearing on his readiness for reading. Children who have been talked to and who have talked much themselves, who have had stories read to them, who have possessed picture books, and who have seen their parents read for pleasure tend to have an initial advantage over those who have had less of these experiences. Conferences with parents or home visits can provide some information. In addition, the pupils themselves will talk about their

⁴ Extensive discussion of visual, auditory, and other sensory measurement is given in school health textbooks.

experience at home if given the chance. In using any pupil reports, careful attention must be given to the six-year-old's penchant for inaccuracy, exaggeration, and sheer invention.

Readiness Tests as a Measuring Procedure. Reading readiness tests are the only systematic procedure many teachers use to measure readiness. Such tests can be effective in appraising phases of a child's language development, his information, and certain aspects of his mental development; and, as we have seen, these are dimensions of reading readiness. However, readiness tests are *tests*. For most pupils they may be the first experience with a test and the strangeness alone may depress many pupils' scores. They require clear understanding of and strict adherence to directions and a child who is ready to learn to read still may be unable to understand or to follow complicated directions the first time he is exposed to them. The tests are necessarily short and thus their sampling of dimensions is limited. Therefore, primary teachers are urged *not to depend* on a readiness test score as the *only index of readiness*.⁵

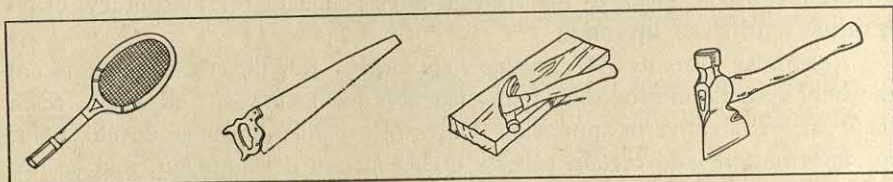
Used along with other procedures, readiness tests are an invaluable aid to the first-grade teacher in gauging the readiness of pupils for reading. Many reputable ones are on the market. They are relatively inexpensive, easy to administer, easy to score, and manuals explain the significance of raw scores and usually give tables for converting them to percentiles or other derived scores. A number of items from two tests are presented in Figure 40 to illustrate the nature of readiness tests.

As a practical consideration, a school or school district may need to choose between administering an intelligence test or a readiness test in the first grade. While no hard and fast rule seems advisable, it is thought that the intelligence test should be given priority because it has significance for all aspects of the first-grade program. *Moreover, scores on mental tests used alone are about as predictive of success in reading as are scores on readiness tests used alone.*

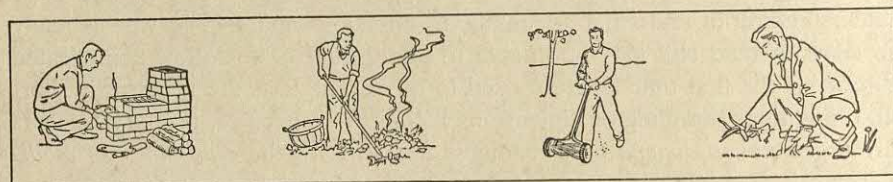
STANDARDS OF READING READINESS

At the outset of this section on measuring readiness, it was stated that a child with a mental age of six years, six months, and with all the normal concomitants of this level of mental development, was considered to be ready

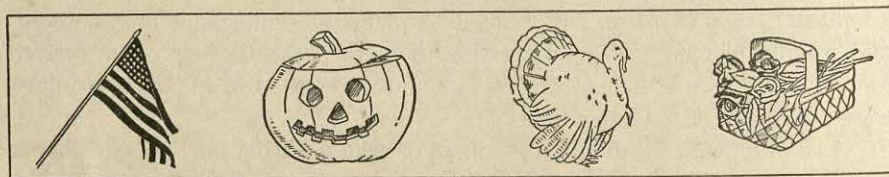
⁵ Designers of readiness tests standardize them on primary pupils and thus norms are somewhat adjusted to the presence of just such variables. This, however, assures their validity more for groups than for individuals and it does not increase their reliability. It is interesting to note that readiness tests and teachers' forecasts of readiness (based, presumably, upon casual observation of and reflection upon many of the dimensions of reading readiness) may not be too closely correlated and hence neither are infallible predictors of success in reading. For example, one researcher, Henig (14), found the equivalent of a coefficient of correlation of .60 between teachers' ratings and scores on the Lee-Clark readiness test administered to 98 pupils. The test bore a correlation of .55 with marks given in reading at the end of the year, while teachers' forecasts showed a coefficient of .59 with end-of-the-year marks.



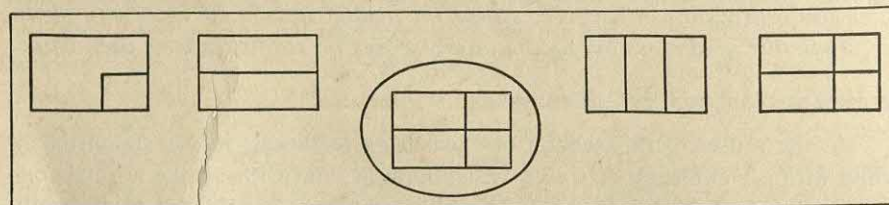
Teacher says, "Mark the hatchet."



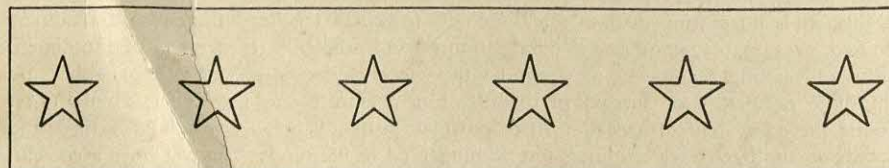
Teacher says, "Draw a cross on the right picture. In the Fall, father rakes the leaves and burns the leaves."



Teacher says, "Mark the one to carry in the parade on the Fourth of July."



Teacher says, "Draw a frame around the picture that is just like the one in the middle."



Teacher says, "See the stars. Mark four of the stars."

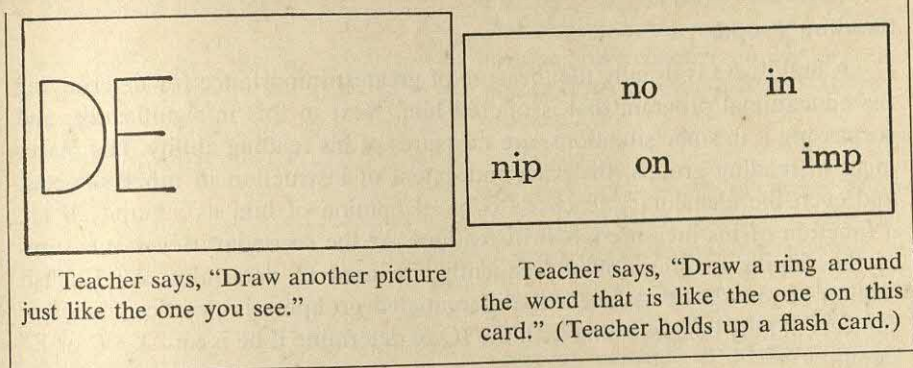


Figure 40. Sample items from readiness tests. (The first six are from the Metropolitan Readiness Tests, Form R, Kg-1, 1949; the last from the Murphy-Durrell Diagnostic Reading Readiness Test, Primary 1947; both copyright by World Book Co., Yonkers, the publishers, and reproduced by special permission.)

to learn to read. In terms of the dimensions of readiness we have just been discussing, "normal concomitants" mean the sensory development, language development, and experience typical for children whose mental age is six years and six months. By implication, children who are at this level learn to read with average ease; superior ones learn to read with greater ease and speed; and those below the norms have more difficulty or, if they deviate too greatly, are entirely unable to learn to read.

The following outline presents the normal conditions of reading readiness in outline form. Expanded descriptions of typical development for children with MA's of six-six are to be found in the behavior profiles of child development textbooks (see Gesell and Ilg (9) in particular).

Mental development:	MA = 6-6. Scores on language portions of mental test no less than this.
Vision and hearing:	Near-vision normal for chronological age of 6-6 or corrected; no uncorrected astigmatism; history of normal hearing at all voice frequencies.
Language development:	At 50th percentile or better on vocabulary and sentence parts of standardized readiness test; speech at or above median for unselected entering first graders (social and interest aspects as well as vocabulary and usage).
Experience:	History of story reading and book familiarity; experience with the ideas and objects to be encountered in reading program; experience in taking turns, drawing, relating anecdotes.
Readiness tests:	50th percentile or other score which test norms say is minimal for first efforts at reading.

Reading Proper

A child's IQ is usually the measure of greatest importance for determining the educational program that is offered him. Next to this in significance, and surpassing it in some situations, are measures of his reading ability. His placement in reading groups, the pace and extent of instruction in other subjects, and even the elementary teacher's over-all opinion of him as a pupil all are a function of his measured skill in reading. At the secondary level, measures of the pupil's reading ability frequently are used to determine the English section he should be assigned. If differentiated grouping is practiced in other subjects, it may be used along with the IQ to determine if he is an "X, Y, or Z" pupil.

DIMENSIONS OF READING

The dimensions of concern in the measurement of reading are few or many, depending upon the purpose of the measurer. For a quick general appraisal, it probably is sufficient to use *rate* and *general comprehension*. Rate, of course, is a simple and clearly measurable dimension—the speed at which an individual reads expressed, as a rule, in words-per-minute. Comprehension is more complex but, in most cases, it means the accuracy with which a pupil can recall the details of what he has read. Sometimes, to this is added his understanding of the general idea or purport of a passage.

When reading is to be appraised for given subjects and activities, when evaluations are to be made of individuals rather than groups and, in particular, when a pupil's difficulties with reading are to be diagnosed, other dimensions need to be added. Moreover, those of comprehension and rate need to be broken down into more specific components. In the following outline, an attempt is made to list many of the dimensions of reading that are susceptible to measurement. Some of them are the focus of given standardized tests. Others pertain to certain school grades and measurement purposes. It is unlikely that a teacher will be concerned with all of them in any one situation.

Some Dimensions of Reading

Mechanical

Eye movements

Number per line, time and place of fixations, number of regressions; more a laboratory than a school dimension.

Voice and hand accompaniments

Extent of whispering, lip movements, finger reading; of concern in primary grades.

Postural accompaniments

Nature and extent of maladjustive postures; of special concern in elementary grades.

Recognition techniques

Identity of and accuracy of; as, sounding, context, configuration, and syllabification; of concern in elementary grades and in diagnosis of retarded readers.

Vocabulary

Oral

Identity and number of words pupil understands when spoken; of concern in primary grades and in diagnosis.

Sight

Identity and number of words pupil can pronounce and understand when seen; of concern in primary grades and in diagnosis.

General

Identity and number of words known whether spoken or seen; of concern at all grades but particularly in secondary and college.

Rate

Words read per minute assuming good comprehension; should be differentiated according to purpose and material; of concern at all levels.

Comprehension

Sentence

Length and difficulty of sentences pupil can understand, accuracy of same; of concern at all levels.

Paragraph

Length, difficulty, and type of paragraphs pupil can understand; accuracy of same; of concern at all levels.

Passage

Difficulty of passages pupil can understand, accuracy of same without regard to length; of concern at all levels.

Oral reading

Rate and comprehension; in addition, accuracy of pronunciation and word emphasis, appropriateness of tone, rate, and inflection according to nature of material, way of sounding unknown words; of concern in primary grades and at all levels for public speaking and dramatics.

Differential purposes

To skim

To memorize

To follow directions

Rate and accuracy for any of the purposes; of concern in upper elementary, secondary, and college levels.

- To get specific information
- To learn how to do something
- To gain general knowledge of a subject
- To judge the validity of an argument or conclusion
- To judge the literary or technical merit of a passage
- To be entertained or inspired

Differential subjects and materials

- { Mathematics
- { Chemistry
- { Etc.

- { Advertising
- { Maps
- { Charts and graphs
- { Etc.

- { Fiction
- { Poetry
- { Etc.

Vocabulary, rate, and/or comprehension for any subject or material; of concern in upper elementary, secondary, and college levels.

Special reading tasks

Rate and accuracy in use of table of contents, indexes, glossaries, directories, dictionaries, encyclopedias, footnotes, etc.; of concern from intermediate grades on.

Special disabilities

Identity and extent of systematic errors, letters and common words not known, etc.; of concern in diagnosing cases of reading retardation.

PROCEDURES AND FORMS FOR MEASURING READING

Of the possible methods of behavioral measurement, only product analysis is not readily adaptable to appraising a pupil's status in reading. Observation is necessary for the mechanical dimensions and is valuable for the several special reading tasks of index and dictionary usage, etc. Often, the most valid approach to measuring comprehension is a free-response procedure—the pupil summarizes or interprets what he has read in writing or orally. Guided response instruments, however, are employed more than any other device and can provide reasonably valid and highly reliable measures of most dimensions of reading.

Reading Tests. Various types of guided response items used in reading are illustrated in Figure 41. The fact that the samples are adapted from

Rate

17. Several of the objects in the night sky are not
18. stars but planets.
19. One of the most interesting planets is the
20. ringed one called Saturn. Its rings are formed
21. by a cloud of small particles that circle the

(Pupil checks the number of the line he was reading when the examiner called "stop.")

*Comprehension**Sentences*

So intense were the flames from the factory that it was truly a *conflagration*.

- (a) steel mill (b) big fire (c) celebration (d) false alarm

(Pupil indicates the best synonym for *conflagration*.)

Paragraphs

John, who weighs 150 pounds, wants to fight in a class which has a top weight of 135 pounds. He intends to lose ten pounds and this will make him eligible.

- (a) eligible (b) intends (c) ten (d) lose

(Pupil indicates which of the option words "spoils" the meaning of the last sentence.)

Long passages

(Pupil reads a given passage—timed, usually—and answers a series of multiple-choice questions about what he has read.)

*Differential Purposes**General significance*

11. Europe is the home of the white storks. People are delighted when they fly north in the spring. Many believe that storks bring good luck to a village. A family is proud when storks choose their roof on which to build a nest. Storks do bring one kind of real luck—they eat anything that is thrown out. This helps to keep the village clean and healthy.

Draw a line under the word that tells what many people believe storks bring.

food nests spring riches luck

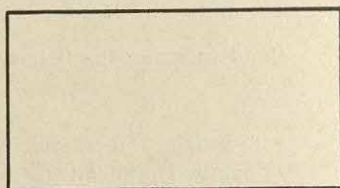
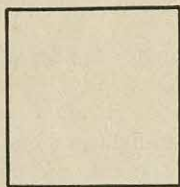
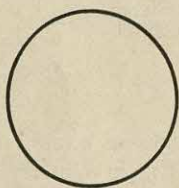
Figure 41. Types of guided response items used to measure various dimensions of reading ability. (The four items for "differential purposes" are by permission from *Gates' Basic Reading Test*, for Grade III (second half) through Grade VIII, Teachers' College, Bureau of Publications. The balance are adapted from those appearing in many tests.)

Predicting outcomes

7. No other creature of the woods is more timid than the deer, yet when winters are particularly severe, deer have been known to visit farms and to accept food. One February after a stretch of very cold weather we saw deer tracks around our barn. That night we threw some hay outside and along with it some kitchen scraps. In the morning we went outdoors.

Our horses were eating the scraps.
 We saw two deer eating the hay.
 We saw fox tracks in the snow.
 We saw the deer pulling our sleigh.

(Pupil underlines one of the sentences.)

Understanding precise directions

22. These different shapes all have different names. The first is a circle; it has no corners. The second is a square; it has four corners. Do you know the name of the third? It is longer than a square. Make a cross in the center of the one that has no corners.

Noting details

A book written nearly three hundred years ago tells a story about a tree which grew in America. The story said that this tree cried when it was cut. It also said that tears came from the cut which dried into a sweet sugar. Now we know that this crying tree was the sugar maple, and that the sweet tears became maple sugar when they dried.

An old book tells a story about a

flag tree book man

The story said that if this tree was cut it would

laugh cry sigh eat

The sweet tears from the tree become

salt powder sugar flour

(Pupil underlines one word for each of the questions.)

Figure 41. (Continued)

Vocabulary

Option (a) engine (b) sweet (c) hurry (d) choice

(Pupil selects synonym from options.)

*Special Reading Tasks**Index*

On one page a portion of an index. On the other, such multiple-choice questions as:

On what page will you find a picture of an albatross?

(a) 37 (b) 99 (c) 138 (d) 22

Maps, tables, figures

On one page a map and a population chart for cities. On the other, such multiple-choice questions as:

The city with the smallest

(a) London (b) Berlin

is nearest on the map to

(d) Hamburg

Directory

On one page an office directory with multiple-choice items associating names

David D. Jones

(a) 107 (b) 214 (c) 318

is building. On the other, multiple-choice questions, as:

1

Figure

used)

standardized tests suggests, as it is the mainstay of measurement in education. Surprisingly homogeneous so far as reading materials and the pace of learning throughout the country, the growth of reading tests seems entirely legitimate. The teacher from devising and using his own items like those in Figure 41 to

The raw scores on standardized tests are usually converted into grade equivalents and/or percentiles. These are indexes of rank and also tell where a pupil stands within the large group of pupils to whom the test was given. A given reading grade means that the pupil's score is the same as the average score made by pupils in that grade. For example, if Bill scores a reading grade of 6, this means

⁶ A number of reading tests together with other standardized tests.

standardized tests are currently used because the United States is so large and because the needs of instruction are much the same everywhere. The use of teachers on standardized tests, however, this should not deter any use of response instruments, adapted to the purposes.

Tests usually are converted into grade equivalents. These, as you know, tell where a pupil stands within the large group of pupils in the standardization process. A grade of 6 means the same as the average score made by pupils in that grade. For example, if Bill scores a reading grade of 6, this means

his remarks are listed in Appendix

that the average score of sixth-grade pupils, in the standardization population, was 63. If the reading grade is fractional, say 6.7, it means that the pupil's reading ability supposedly corresponds to that of an average pupil who has been in the sixth grade for seven months. A percentile score on a reading test has significance only within a specified grade or grade segment and should not be used unless the group to which it refers is indicated or clearly understood.

Scores on tests devised locally should be considered as indicative of rank only and may be converted into percentile scores (see page 157). If raw scores on the test have a restricted range or if only gross differentiation among pupils is necessary, the determination of quartile limits (see page 152) and thus the classification of pupils into upper, lower, and intermediate quarters may be all that is necessary. Any free-response instrument used to assess reading may yield classification or rank scores. Measures derived from observations of reading must, as a rule, be verbal descriptions or classifications according to some rating system. *Scores derived from any nonstandardized procedure should not be converted into reading grade equivalents.*

It is becoming ever more common to express competence in reading as a series of scores or as a graph showing the status of the pupil with respect to the several dimensions that the test measures. Since reading is a complex of many dimensions, this profile approach, as it is called, seems to be a far more appropriate way to measure a pupil's reading ability than is a single score. See pages 125, 373 for illustrations of test profiles.

Cautions in Use of Standardized Tests. In using standardized tests certain cautions must be observed. It is probable that they will have limited validity for the extremes of norm groups. For example, a test designed for grades 4, 5, 6, 7, and 8 is less valid for fourth and eighth graders than it is for grades 5, 6, and 7. Stanley (24) found that the *Nelson-Denny Reading Test* for High School and College was too difficult for the lower half of a typical ninth grade. Secondly, nearly all standardized tests are *timed* tests. In sections that purport to measure comprehension, the pupil is likely to try to read as fast as he can and his rate may be greater than his usual rate. Consequently, for many pupils it is probable that *comprehension during acceleration* is being measured and not general comprehension (21). A third point of caution is that the vocabulary, content, and format of the reading tasks of standardized tests may be somewhat unlike that of the material the pupils normally read. These differences will have a particularly depressing effect on readers with little self-confidence, those who show appreciable test anxiety and those whose reading is nearly all school-connected.

Diagnosing Disabilities. The diagnosis of reading disabilities is an important but highly complex undertaking. The use of standardized diagnostic instruments is time-consuming and requires a more extensive knowledge both of measurement and of reading than may be presumed for the average teacher. Descriptions of diagnostic procedures and instruments are to be found in

textbooks on reading method and on remedial techniques. Diagnostic dimensions are essentially no different from those presented in the outline on pages 228-230. They are, however, likely to be greater in number and more detailed and individualized than those measured by ordinary reading tests.

A nonstandardized observational procedure, recommended by Dolch (4), can yield significant, if imprecise, information about a poor reader's difficulties and requires no special training for its use. In this procedure, a pupil first reads a passage aloud and the teacher supplies any words over which he hesitates. This can disclose the *common words* with which he is unfamiliar. Second, he reads aloud another passage, is given no help, and then is asked to relate what he has read. This should afford some evidence of his *general comprehension* of what he reads. A third passage then is read by the pupil. Each time he hesitates over a word, he is asked to guess what it means and thus a measure is provided of his ability to use *context clues*. Finally, in a fourth passage, he attempts to pronounce and tell the meaning of unknown words in order to determine what *word attack methods* he uses and how skillfully.

EVALUATIVE STANDARDS IN READING

The only precise objective standards by which a pupil's reading ability may be evaluated are the grade norms of standardized tests. These indicate the usual range of reading test scores for given grades and, presumably, the place of a pupil's score in this range is indicative of how bad to fair to good is his reading ability for that grade. In addition, a pupil's reading grade placement as derived from the norms of a standardized test may be compared with his actual grade placement. If the reading grade is higher than the actual grade, his reading ability may be judged favorably; if lower, adversely.

The relationship between given rates of reading and absolute indexes of comprehension, on the one hand, and degrees of success in school and occupations on the other, might provide a nonrelative and independent standard. Unfortunately, although the relationship is known to be positive and appreciable, it has been determined with no precision. So far as rate is concerned, it is possible to use the word-per-minute equivalents of grade means and percentile ranks in given grades as some sort of nonrelative standard. For example, if on most standardized tests the mean w.p.m. of College freshmen is 250-300, a high school senior whose rate was in excess of this could be judged to be able to read freshmen reading assignments in the time allotted, while one whose rate was less than 250 w.p.m. could be expected to spend undue time on reading assignments.

It is felt that the matter of ability differentials should be given particular consideration in evaluating reading. Surely, immature primary pupils should not be given low marks in reading as a penalty for their immaturity. Yet, if strict competitive grading is practiced, this will result since mental maturity is the most important single determiner of progress in reading in the primary grades. Consequently, it is recommended that descriptive statements, profiles

of status in several reading dimensions, and teacher-parent conferences, one or all, be used for reading in the elementary grades rather than a single letter or word mark.

SPEECH

While effective speaking is a *sine qua non* of education, it is perhaps the least measured of all educational phenomena. There are no standardized instruments listed for its measurement in the most recent Mental Measurements Yearbook. Dimensions are not exactly defined and there are no precise evaluative standards for the speech of school pupils.

Dimensions of Speech

In general, the systematic evaluation of status and improvement in speech is of concern only to teachers of public speaking and drama and to English teachers who include speech or dramatic expression in their classes. In addition, speech correctionists and pathologists engage in the detection and diagnosis of vocal disabilities. However, this sort of clinical measurement is beyond the scope of our text. The dimensions that speech teachers most commonly assess are stated in the following outline.

Rate	Words per minute.
Rhythm	Smoothness—unevenness of talking; appropriateness of pace to subject; observance of meter if poetry.
Pitch	Usually expressed qualitatively as high to medium to low, but can be measured in mean cycles-per-second.
Amplitude	Usually expressed qualitatively as loud to medium to soft, but can be measured in mean decibels.
Tone	Extent of resonance, nasality, harshness, etc.
Enunciation	Correctness and clarity with which words are pronounced.
Diction	Appropriateness and variety of oral vocabulary and usage.
Posture and movement	Manner and appropriateness of standing, facial expressions, gestures, etc.

Several of these might be omitted and others would certainly be added according to the nature of the measurement task. In public speaking, dimensions relating to the content of speeches should be important and in dramatics, such factors as stage movements, ability to memorize, and "stage business."

Forms and Procedures in Measuring Speech

Description or classification and ranking are the measurement forms applicable to most speech dimensions. Rate, of course, may be expressed in mean words-per-minute, a scale number.

Observation currently is the principal measuring procedure appropriate for school use. By the nature of the phenomenon instrumented measurement is likely to have questionable validity. Paper-and-pencil tests must be devoted almost entirely to measuring a pupil's *verbal knowledge* of correct speech, not his *performance* of it; and, unfortunately, there is not a very high or even necessary relationship between the two. It's as though a rifle expert were to have his marksmanship judged, not by shooting at a target, but by answering questions about how one should shoot at a target. In observing a pupil's speech, efficiency will be gained by using an appropriate observation schedule or rating scale. The construction and use of these as well as principles of observation are described in Chapter 4.

In addition to observing a pupil's speech directly, it is possible to record some of it and then listen to the recording. The advantage of this process is that the tape or disk can be replayed and thus the examiner has more time and can listen for specified dimensions without feeling that he is neglecting others. Moreover, a pupil can listen to a recording of his own voice and be directly appraised of his correct and incorrect utterances. Foreign language teachers more and more are using tape recorders both for their evaluations and for student self-evaluation.

Evaluative Standards in Speech

Textbooks on speech and manuals of teaching method represent, in general, the characteristics of effective speaking. Presumably, a teacher may compare how a pupil talks with the book's prescription and judge him accordingly. However, textbooks usually present an ideal way of speech and not those less acceptable, and it is necessary for the teacher to judge how close to the ideal eleventh- or ninth-grade pupils should be expected to come. Moreover, the statements in the book must be translated into visual and auditory sets before the teacher can use them as standards in evaluating a pupil's talk. In this process, the teacher's own style of speaking will have a necessary influence and, thus, the standard actually used by the teacher may be extremely subjective.

A better sort of standard for speech is thought to consist of recordings made by previous pupils that represent the range of competence commonly found in the grade in question. These could be played to the class at appropriate times and the teacher could indicate how he evaluated each specimen, even giving them *A's*, *C's*, and *F's*, if he wished. Pupils then might have a more realistic notion of what they could accomplish and how this might be graded.

As well as furnishing a fairly objective group standard, recordings also can provide each pupil with an individual standard. This would be a recording of each pupil's speech made at the beginning of instruction. At the end or at several times during the instruction, another recording could be made using the same material. A comparison between the two would show how much progress had been made.

Course or subject grades seldom are assigned to speech as such except in public speaking and dramatics classes. Marks tend to be somewhat higher and factors of effort and citizenship may have a greater bearing than in the three R's or in more academic secondary courses. For pertinent procedures of marking and reporting, referral is made to Chapter 9, pages 207-212.

COMPOSITION

Composition is the term usually applied to free writing, where the pupil provides his own words, phrases, and sentences and does not merely copy from a text. It may take the form of exposition, description, narration, or argument and may be imaginative or factual. Elementary teachers usually become interested in evaluating a pupil's ability to compose in the middle grades, say IV or V, and continue their interest through the balance of the elementary grades. In secondary schools only English teachers and, sometimes, instructors in the Social Studies, are much concerned with measuring skill in composition. In our treatment we will concern ourselves only with the general aspects of composition, omitting those unique to its special forms. Evaluative procedures for the latter—"plotting" in narrative writing, rhyme and meter in poetry, for example—are described in student texts and in instructional handbooks devoted to the special forms.

Dimensions of Composition

Probably as many different lists of composition dimensions have been devised as there are English teachers and English texts. But it is thought that those presented in the following outline would be included in nearly all of them. Moreover, the breakdown of the outline is consistent with the breakdown usually observed by standardized tests. As in other language art areas, a teacher seldom will need to measure all these dimensions in a given situation and may wish to add others.

Some Dimensions of Composition

- | | |
|-------------------|---|
| 1. Syntax | Appropriateness and accuracy with verb tenses and forms, plurals, pronoun cases, placement of modifiers, complete sentences, etc. |
| 2. Capitalization | Appropriateness and accuracy with proper nouns, titles, place names, first words in sentences and quotations, etc. |

3. Punctuation
4. Format Indentation, margins, the special forms of letters, verse, outlines, etc.
5. Spelling
6. Vocabulary Accuracy of pupil's definition, appropriateness of words used to ideas, extent of words used accurately, etc.
7. Sentence facility Assuming syntactical correctness, appropriateness of form to thought, parsimony of expression, placement of modifiers, phrases, and clauses, aptness of figures of speech, etc.
8. Paragraph facility Organization (topic sentence, chronology, opposition and resolution, etc.), use of parallel sentence forms, transition words and phrases, coherence and unity of ideas, etc.
9. Style The most ill-defined and variously defined of composition dimensions, includes such variables as variety of expressions, humor, subtlety or obviousness, interest and "warmth"; before attempting to measure style, it must be given operational definition.

The first four dimensions are often called mechanics or grammar and they, along with spelling and vocabulary, easily satisfy our conditions of measurability. They are observable, discrete, well-defined, and agreement as to what constitutes correctness for each of them is relatively easy to obtain. Not so, the last three. These, having to do with rhetoric or effectiveness as against correctness, are observable but they lack the precise definition and agreement among observers, so necessary for accurate measurement. "Style," of course, is the least measurable of all. It is included only because teachers so frequently try to evaluate it and because the attributes it symbolizes, however vague, are of great importance.

Because the rhetorical dimensions are difficult to measure objectively and accurately, it is especially important that they be defined in very specific language. It is not sufficient to say "I want to measure my pupils' ability to write effectively," and then to devise a test or to start judging their written products. It is necessary, as a prelude, to specify the details of these dimensions and the way these details will be manifest in a product or in responses to test questions.

Forms and Procedures for Measuring Composition

Expression of a pupil's status in the dimensions of composition largely is restricted to description-classification and to rank symbols. A number of different systems are used—letter marks, written comments, standard scores

and percentiles on standardized tests, grade placement, vocabulary age, —but none of them has the characteristics of a true scale.

PRODUCT ANALYSIS

Composition is admirably suited to measurement by product analysis procedures. The outlines, summaries, letters, themes, and stories that pupils produce so that they may learn to write are at the same time evidence of their ability to write at any given moment. In fact, for practical purposes a pupil's writing ability is no more than the merit of what he has written, so there is often no need to resort to tests for appraisal of this phenomenon.

In marking written products, a teacher may read the passage and assign a single mark to the passage as a whole. Or, he may use some factor counting or rating system and assign numbers or marks to the several dimensions of the passage. The first method is quick, is consistent with the fact that a piece of writing is a unit and not just a collection of parts; and probably it is the method of product appraisal most widely used by teachers. On the other hand, "over-all impression" marking is less reliable and far more subjective than a detailed analysis and is criticized by most researchers and measurement specialists (27). The College Entrance Board "Essays" are marked differentially for five dimensions and the Educational Testing Service scores its Foreign Service essay examinations in similar detailed fashion. It is thought that the increased reliability and diagnostic value of a detailed analysis more than justify the extra time it requires. Techniques for this type of product appraisal are explained in Chapter 5, pages 76–81.

COMPOSITION TESTS

When it is desirable to measure achievement in composition by a test of some sort, the whole or part question asserts itself in another way. It is possible to ask for a test response that constitutes an act of writing. It is equally possible to ask for responses that are not in themselves the actions of writing but bear an essential relationship to writing. An example of the first type of test is the College Entrance Board *General Composition Test*, Form F, which provides specific reading material on which an essay is to be based and then requires that the student write an essay of several hundred words on this material. The second test procedure is exemplified by true-false and multiple-choice questions about rules of grammar and by items that ask for the meaning or spelling of words.

You may recall the statement in Chapter 3, page 41, to the effect that maximum validity may be expected for a behavioral measuring instrument when it instigates the actual phenomenon subject to measurement. According to this viewpoint, tests in which pupils write or somehow do the very things persons do when they write, have the greater potential validity. Drawbacks to this type of test are that it may be inefficient in use of pupil time, the reliability of scoring may be low and, in a given testing period, only one or a few of the varied forms and situations of writing may be sampled.

The other approach—to elicit responses that are not themselves the usual acts of writing—must base its claim of validity on the premise that the test responses are directly *related* to how a pupil writes. While this is recognized by psychometricians as a legitimate rationale, the relationship must be demonstrated and the relationship often is too small to assure an adequate degree of validity in the test. The advantages of such an indirect measuring procedure lie in its speed of administration, in the reliability and even automation of its scoring, and in the careful and comprehensive sampling of the various forms and situations of writing that can be accomplished in a single testing period.

Direct and Holistic Procedures. There seem to be about three possible ways of directly testing a pupil's composition ability:

1. Rating or analyzing a passage written on demand,
2. Comparing such a passage with a product scale, and
3. Editing a presented passage.

The first is well illustrated by the College Entrance Examination just cited. The dimensions with which this standardized test is concerned and the manner of rating them is shown in Figure 42, a reproduction of the grading sheet to be used by the examiner for each essay. Ratings mean: 1, superior, 2, high

GRADING SHEET					
NO. OF PAPER _____			NO. OF READER _____		
DATE _____		TIME BEGUN _____		TIME FINISHED _____	
1	MECHANICS	STYLE	ORGANIZATION	REASONING	CONTENT
Border					
2					
Border					
3					
Border					
4					

Place papers in categories 1 to 4 with an X; border may be marked only $\wedge$ or $\vee$.

Figure 42. Grading Sheet from the College Entrance Examination Board, General Composition Test, Form F, Educational Testing Service, Princeton. Reprinted by Special permission of Educational Testing Service.

adequate, 3, low adequate, 4, inadequate. As may be apparent, the central weakness in this and any other essay rating procedure is the inability of examiners to agree exactly as to their working definitions of dimensions and rating categories and even to achieve complete self-agreement on reappraisal of the same paper. The aspects of the composition that may be measured most reliably are those that involve specific errors and can be counted: syntax, punctuation, and spelling.

The use of product scales, the second procedure, is infrequent possibly because so few reliable and well-standardized scales have been published. Buros cites only *Hudelson's Typical Composition Ability Scale*. According to the reviewer, Osburn (20:316-317), it is the best product scale extant but still is an imperfect instrument. The scale consists of a series of paragraphs varying in excellence from that representative of the fourth grade to that of the twelfth grade. A teacher finds the point of best fit and assigns a pupil the number corresponding to that point much as is done in rating handwriting specimens (see page 252). Other product scales are the *Hillegas Composition Scale* (the forerunner of all such scales), the *Willing Scale for Measuring Written Composition*, and the *Lewis English Composition Scales*.

The third holistic procedure, the copy-reading test, is illustrated by the third sample item in Figure 43. Copy-reading tests can be highly reliable, and, for this reason, are widely used in standardized English batteries. Their validity, however, is suspect. The technique requires the recognition and correction of existing errors. This, of course, is something a writer may do in rereading what he has written but *it is not what he does when he composes originally*. So the greater part of the validity of the approach must rest on an assumption of direct and invariable relationship between skill in original composition and skill in editing. Such test investigators as Travers (27) and Griffin and Venable (13) think that this relationship has been exaggerated and, for that reason, question claims of high validity for the copy-reading procedure.

Indirect and Atomistic Procedures. As we have indicated, a second basic way to test a pupil's composition ability is to appraise separately each of the many items of knowledge and skill that presumably agglomerate to form his writing ability. The great majority of published tests use this approach. There are composition or English "batteries" which, in a single testing period, test a sample of all the important knowledge and skills involved in writing and there are as well separate tests for each basic subdivision of skill or knowledge. The usual components of batteries and the subjects of separate tests are mechanics (syntax, punctuation, etc.), spelling, vocabulary, and rhetorical effectiveness. Guided response items used for the measurement of mechanics are illustrated in Figure 43 following. Spelling items are shown in Figure 44, page 245, and those testing rhetoric in Figure 45, page 247. Figure 41 in the section on reading also contains a sample vocabulary item. The types of items in the figures, while derived from standardized tests, are just

5. Let them stay a little longer if they want to; let's us go home.

(Pupil selects the underlined part he thinks is wrong.)

11. There ^(11-1 wasn't)_(11-2 weren't) many ships like the one on which my grandfather sailed long ago.

(Pupil selects the correct verb form.)

John walked to the door with

his guest. "let me know when

20

20 ()

20
you are in town again, colonel

21

21 ()

johnson," he said. "if I am not. . .

22

23

22 ()

23 ()

(Pupil writes *c* in the () if he thinks a word should be capitalized, *s* if he thinks it should start with a small letter.)

. . . I seem to be one of those fortunate people who recognizes that its a privilege to see the common place beauties which surround us. Why don't everyone see them? Hasn't it occurred to. . .

(Pupil copyreads the passage and edits it as he thinks necessary.)

I wonder what hes planning to do whispered

44

45

46

Frank to John I do not think he has any real injury

47

48

do you John thought it best to make no reply

49

50

51

(Pupil indicates what punctuation, if any, is needed above each of the numbers.)

Figure 43. Examples of guided response items designed to test a pupil's knowledge of the mechanics of composition. (All are from the Cooperative English Tests for High School, published by Educational Testing Service, Princeton. Reprinted by special permission of Educational Testing Service.)

as appropriate for nonstandardized instruments. It is interesting to note that the items depart from the traditional true-false, multiple-choice types, thus demonstrating the flexibility of guided response procedures.

Mechanics. The measurement of a pupil's knowledge of syntax, punctuation, etc., as things apart from his actual writing is open to several criticisms. Correct usage does not invariably follow knowledge of the rules and elements of correct usage. Manuals and texts covering mechanics often differ as to what is correct usage. The inclusion of items in tests often seems to be

based on frequency or importance in English texts rather than in public writings. These, and other points of criticism, are supported by research but the measurement of mechanics by means of guided response instruments persists. The analysis by Griffin and Venable of eleven standardized tests of usage (13) is a representative critique of this aspect of compositional measurement.

In our view, analysis of the mechanical errors in whatever products the pupils write is a far more valid and sufficiently reliable means of measurement for mechanics. If a test is desirable, passages can be written on topics designed to elicit certain usages or pupils can be instructed to write sentences and/or paragraphs containing the elements a teacher wishes to appraise. For example, "Write a sentence containing direct address and another containing indirect address." "Compose three sentences about a boy, a shotgun, a dog, and a rabbit. One should be a simple sentence, one a complex sentence, and the third a compound sentence." It is necessary to resort to standardized tests and thus to guided response items for mechanics only when comparative data is sought, i.e., the grade placement of a pupil's knowledge of mechanics.

Spelling. A pupil's ability to spell may be judged in either of two ways. The writings he submits may be examined for misspelled words with frequency of error considered to be an inverse index of his spelling ability. The chief limitation of this method is that any pupil's free-writing vocabulary may be unduly restricted to words that he can spell. Its primary advantage is that spelling is measured as it is ordinarily done.

The second way is to ask pupils to spell a number of specified words. This method has the advantages that attend use of a controlled list of words. They can be made easy or difficult at will. They can be chosen to represent given types of words. If the list is extensive and carefully chosen, it can be considered a representative sample of all words. Moreover, pupils can be compared with each other more fairly than when product analysis is used. As its disadvantage, the spelling list method gauges the spelling of words in isolation and under conditions that exist for the test only.

It would seem that both methods are necessary if spelling ability is to be measured efficiently. The first provides the more diagnostic information and can be performed casually and continually. The second is essential for comparative data and for any comprehensive appraisal of the words pupils can and can't spell.

The first task in devising a spelling test is to determine the words to be included. It is common practice in elementary grades today to select these from among the words pupils use in school and/or are exposed to in their readers, subject texts, and collateral materials. If spelling is taught as a specific subject and if spelling lists are used for the instruction, the words on the lists are the ones to be included in the tests.

The words to be used in any test should be judged in terms of appropriate difficulty and adequate sampling. Publishers of standardized spelling tests

generally grade the difficulty of words they use for any grade level from those most pupils can spell to those very few can spell. This is a good criterion for a teacher to follow in devising his own tests. If the speller lists words in order of predicted difficulty or according to the grade in which they should be learned, it is relatively easy to make a selection.⁷ It is not expected nor necessary in teacher-devised tests that the difficulty of words be arranged on a precise statistical basis as is done in some standardized tests.

How many words should be included on a given test can not be stated generally. The more words, the more reliable are likely to be the results. Semester or year tests ordinarily should be longer since more attention may be given to scores on them. (See page 103 for a general discussion of the number of test items necessary for adequate sampling.)

Spelling tests may be administered orally or in writing. If the administration is oral, words should be pronounced, used in a sentence, and repronounced. Types of items used in written spelling tests are shown in Figure 44.

1. Check the word which is misspelled
— liability — ocasion — deficit — none of these
2. The correct spelling of a word meaning a military officer is
(a) Lutenant (b) Lieutenant (c) Lieutenat
(d) Liutenent (e) none of these
3. Underline the misspelled words in this passage
... There are many things to obsorve as you ride down town on the bus.
Besides people and animuls, there are stores, cars and constrution jobs. ...
4. The examiner pronounces words for pupils to spell.

Figure 44. Examples of guided response items and techniques for measuring spelling ability.

The selection of a correct spelling among several incorrect ones for a given word, Sample 2, is perhaps the type most frequently used in published tests.

Standardized spelling tests are available, usually as parts of batteries. Whether to use a standardized test or a locally devised one depends largely upon the purpose of measurement. If this is to evaluate achievement during a given period of instruction, a test devised from the reading material and spelling words taught would seem to be appropriate. If the purpose is to compare a pupil or a class with age or grade norms, then the standardized instrument is mandatory.

⁷ We are talking here about measuring general achievement in spelling. If the purpose is merely to see that pupils have learned the words they were directed to study, these are, of course, the items for the test.

It sometimes is desirable to appraise not only how well or how poorly a pupil spells but why he spells poorly and what words he has trouble with. Among the diagnostic tests available to provide this information is the *Gates-Russell Spelling Diagnosis Tests* (Appendix B., page 470). This instrument permits analysis of a pupil's phonetic and sight disabilities, his method of studying spelling words, his auditory discrimination, etc. Examination of a pupil's misspellings in any written work he submits can also be a diagnostic procedure. Apparent in his writings will be the words he chronically misspells, his transpositions and reversals, his phonetic confusions, etc.

Vocabulary. As with spelling, a pupil's vocabulary may be assessed by examining his written products, to see what words he uses correctly and incorrectly, how varied are these words, and to what subjects and abstraction levels they belong. However, this is a seldom-used procedure. When separate measurement is given to vocabulary, it largely is through the device of a guided response test and usually this is a standardized one. Nearly every English and general achievement battery contains a vocabulary section and both group and individual intelligence tests measure vocabulary as a significant component of intelligence.

In interpreting the results of vocabulary tests it is well to keep certain things in mind. Words are included because of their relative frequency of occurrence in adult writings or in school use, much as with spelling. In the tests, words usually are graded as to difficulty on the basis of what percentage of the standardization population knew them. Their difficulty in this case has nothing to do with the complexity or abstractness of the idea they symbolize. The words in standardized tests have a strong middle-class and literary bias. Finally, the tests, as a rule, do not measure the comprehensiveness or precision of a pupil's definition but only his knowledge or lack of knowledge of a given minimum and rudimentary definition.

Rhetorical Effectiveness. The measurement of such dimensions as sentence and paragraph facility and style seldom is undertaken separately and apart from actual writing. When guided response procedures are used, they usually require discrimination between good and bad phraseology and/or knowledge of principles of good writing. Two illustrative items are presented in Figure 45.

One of the criticisms of these procedures is much the same as the one voiced for guided response tests of writing mechanics: they test skill at editing, not writing, and the relationship between ability to edit and ability to write is known to be imperfect. Another point of criticism is equally telling. What constitutes good and bad phrasing is sometimes debatable and, hence, what are the correct answers may be indeterminate.

The effectiveness of any piece of writing ultimately depends upon whether a reader considers it effective, whether he responds to it as the writer intended. Only a limited number of "rules" may be derived for such "effective" writing. Moreover, any student of language knows that these rules change from region

- 13-1. The creek overflowed its banks due to heavy rainfall.
 13-2. The creek overflowed its banks, heavy rainfall being the cause.
 13-3. The overflowing of its banks by the creek was owing to heavy rainfall.
 13-4. The creek overflowed its banks because of heavy rainfall.

(Pupil selects the best sentence of the four)

<i>Column 1</i>		<i>Column 2</i>	
A-1	{ Housing seems to be as much of a problem in the bird world as it is among human beings.	A-2	{ Housing, a problem in the bird world, is also quite a serious problem among human beings.
B-1	{ A one-family bird house was inspected by two pairs of bluebirds in a suburban garden the other day.	B-2	{ The other day two pairs of bluebirds inspected a one-family bird house in a suburban garden.
C-1	{ Both pairs liked it. Both decided to move in. But did they draw straws for priority? They did not!	C-2	{ Both pairs liked it and they decided to move in, and they did it draw straws for priority no indeed.
D-1	{ The males haggled a bit, got nowhere, then flew at each other, beak and claw, and fought it out.	D-2	{ The males haggled a bit. They got nowhere. They flew at each other beak and claw. The matter was then fought out.
E-1	{ The winner of the battle and his mate moving into the bird house, quite happy in their new home.	E-2	{ The winner of the battle and his mate moved into the bird house and are now quite happy in their new home.

(Pupil reads each of the columns and then answers questions which require that he select the better phrased passages and know why they are better)

A. Section A is better expressed in

A-1 Column 1.

A-2 Column 2 A ()

a. The inferior version of Section A is poor because

a-1 emphasis is placed on the wrong part of the idea.

a-2 the sentence is incomplete.

a-3 two sentences are punctuated as if they were a single sentence.

a-4 it is grammatically incorrect a ()

Etc.

Figure 45. Two examples of guided response items keyed to rhetorical skill. (Both are from Cooperative English Tests for High Schools, Educational Testing Service, Princeton. Reprinted by special permission of Educational Testing Service.)

to region, from culture segment to culture segment, and from year to year. For these reasons there are few standardized English tests that attempt to measure rhetorical effectiveness through guided response items. The Co-operative English Tests (Appendix B, page 470) and some forms of the College Entrance Examination Board English Tests are perhaps the more notable examples of such tests.

One alternative to guided response measurement is analysis of products or test essays in terms of rhetorical dimensions. This is a phase of the whole method of measuring composition previously described (see page 240) and is the procedure most used by English teachers. A second alternative is available for certain phases of the rhetorical dimensions. It consists of providing pupils with given words, facts, or ideas and asking for their expression in certain forms. For example:

(1) Write a sentence containing a figure of speech that would describe a *teacher scolding a pupil for misbehavior*.

(2) Rewrite these three sentences to give them parallel construction and put them in proper sequence to make the best sense:

The rivers on the plains were swollen by water from the creeks. Down from the storm clouds in the mountains came water to fill the creeks. A tide of water from the rivers surged out into the sea.

The virtue of this technique is that it requires writing, not just editing or recognition of propriety, and it requires the same sort of expression from all pupils, thus permitting some comparison among them. Such items may not, though, be scored as objectively and reliably as strictly guided response items and they may be applied only to recognized rhetorical conventions.

Evaluative Standards and Practices in Composition

As in the majority of school subjects, teachers use largely subjective standards for judging the value of their pupils' achievement in composition. These subjective standards, what teachers feel is good, fair, or bad work at any grade, seem to derive from two objective sources. One consists of the English texts, the dictionaries and, above all, the published writings that the teachers have read and admired. The second is the writing typical of pupils in the grade levels in question. As suggested in Figure 46, the influence of classics and style manuals seems to set one axis of the standard and the influence of all the pupil themes and letters ever read seems to set the other axis. Subjectivity enters because the two influences or axes are never the same for two teachers and because the relation of *A's*, *C's* and *F's* to the two axes is partially a function of the temper and training of the given teacher.

The use of such two-dimensional standards for evaluating composition seems to be valid. The published writing accepted as exemplary by educated adults is, by definition, the basic standard for adult writing and should affect school standards. On the other hand, young pupils should not be expected to

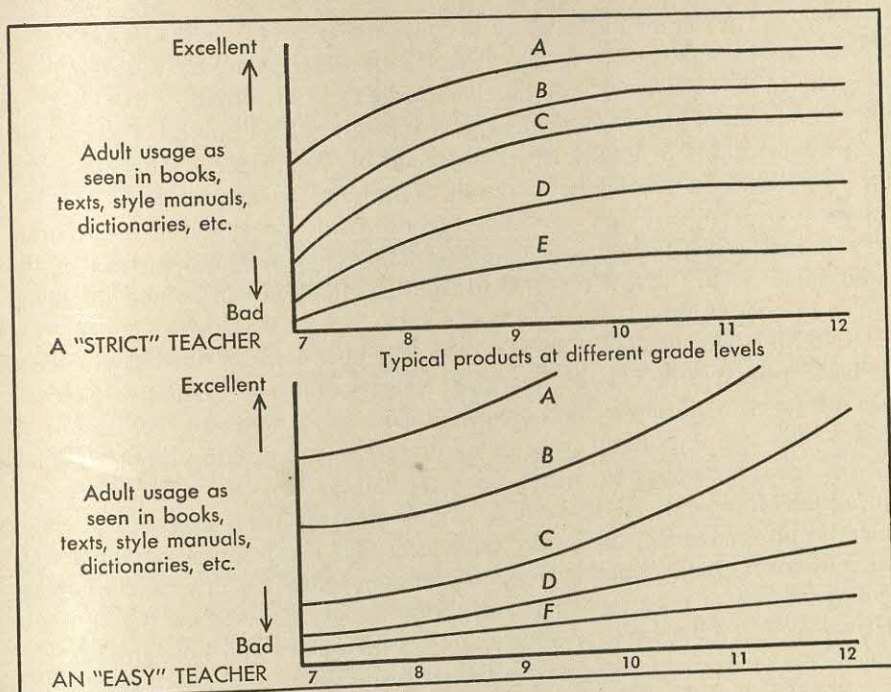


Figure 46. How evaluative standards seem to operate for English composition.

write as well as older ones, so grade differentials in writing should affect school standards as well. It is thought that adherence to the following rules will do much to reduce the subjectivity and capriciousness of composition standards.

1. Write down as explicitly as possible the type of writing and the degree of correctness that will be assigned each mark at each grade level.
2. Collect many samples of pupil composition awarded different marks according to the standard.
3. If standards are to be different for different ability levels, make this difference explicit in the standards.
4. Within a school and within a school system, the statement of standards and collection of sample pupil products should be a group undertaking. Insofar as possible, the same standard should be used by all teachers handling comparable pupils in comparable classes.

Standards for spelling involve a special consideration. As we have observed, the words children must learn to spell in elementary grades usually are controlled by a speller or by a list of some sort. With such an objective and exact instructional objective, the standard for evaluating spelling achievement can also be objective and exact. Correct spelling of the words accumulated up to a given grade might be set as the equivalent of a given mark, with

greater and lesser amounts being assigned other given marks. The proper words/marks equivalence would need to be determined by a teacher according to his experience and the characteristics of his pupils.

Composition seldom is marked as a separate school subject. On elementary report cards it usually is a component of "language," "language arts," or "English." In secondary grades it contributes to the grade in English classes. Only a very few secondary report cards provide for its separate evaluation. When conferences or anecdotal reports are used instead of the traditional *A, B, C, D, F* report card, specific attention can be and usually is given to composition.

Not to evaluate composition as a separate phenomenon obviously violates a basic principle of efficient marking: the meaning of evaluative symbols should be clear to pupils and parents. When a single letter or word refers to more than one important phase of a subject, as to speaking, understanding of literature, composition, handwriting, etc., there can be no clear communication. Moreover, just to evaluate composition by itself may not be sufficiently informative to pupils. We have seen that composition has numerous dimensions. Unless a pupil achieves to the same extent exactly in all of them, a single mark or even an evaluative phrase is only an average. The pupil is left to surmise how weak, relatively, are his mechanics and how strong, relatively, his vocabulary and style, in the teacher's view.

So, it is recommended that composition be evaluated as a separate subject and that it be evaluated with respect to each of its important dimensions. This may be done by conversation, by written statement, by letter or number marks for each dimension, by a profile graph or by a combination of all. The use of profiles has been discussed in connection with measuring procedures (see pages 125-126), but the profile is equally applicable to reporting evaluations. A profile report card for English, including composition and its dimensions as separate entities, is shown in Figure 49.

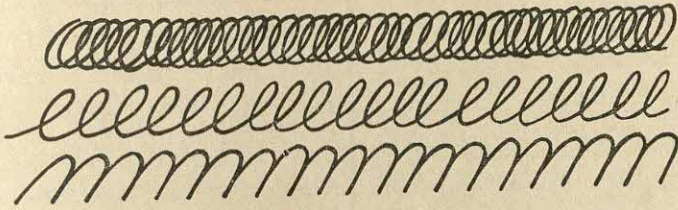
HANDWRITING

In this day of the typewriter, excellence of handwriting is rapidly becoming a lost art. Even the smallest businessman uses the typewriter for billing and business correspondence. The official records of court proceedings, land ownership, marriage and death, now are found neatly typed rather than elegantly inscribed as formerly. The use of handwriting still persists, of course, for notes, for personal correspondence, and above all, for the themes, outlines, tests, and homework of elementary and secondary schools. In the schools, consequently, ability to write legibly and attractively is an important asset still but in the adult world it seems no longer to be.⁸

As handwriting has diminished in importance, instruction in handwriting

⁸ There seems to be no conclusive research on this point but most educators seem to think that handwriting is socially less important today than yesterday.

has received less stress and evaluation of achievement in handwriting has languished even more. In the 1910's and 1920's Ayres, Thorndike, Freeman, and others with their "normative" scales and diagnostic procedures seemed to offer elementary teachers scientific ways of assessing progress in handwriting. As teachers taught pupils the correct slant, the full arm movement, and had them practice



so did they compare the pupils' writings with the scales, give the pupils timed tests, and try to analyze their difficulties. In the 1950's, however, drills in O's, l's and m's have become far less extensive. Manuscript writing largely has replaced cursive writing in the primary grades, and the Ayres, Thorndike, and Freeman scales are used much less frequently *though they still are the standard instruments for measurement of handwriting*. That no important new standardized instruments have been devised for handwriting since 1933 is a telling indication of declining interest in its measurement.

Handwriting Dimensions

Legibility, quality, and speed are the dimensions of handwriting most frequently "measured." Obviously, the first two dimensions are a function of the observer as well as the writer. Attempts have been made to objectify legibility by analyzing it into such components as letter formation, spacing, slant, letter height, lightness—heaviness of stroke and regularity. To make the dimension of quality (beauty, attractiveness, etc.) less subjective has been far more difficult and the effort has met with little success.

Means and Forms of Measurement for Handwriting

The most reliable way of measuring handwriting, through use of a product scale,⁹ begs the question of dimensions. A sample of a pupil's writing is compared with scale specimens until one most like the pupil's is found. The number or other index of value belonging to this scale specimen is the measure

⁹ While these instruments are called scales and while their designers have attempted to give them the attributes of true scales, it is thought best *not* to treat scores derived from them as scale numbers but simply as rank numbers. The different grades of handwriting in a scale necessarily represent rank order among the specimens considered for the scale. It seems to be more reasonable to consider merely that the "scale" is representative of rank order for a given group of pupils than to consider that it is representative of regular gradations of skill in handwriting.

60	70
<i>Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal</i>	<i>Four score and seven years ago our fathers brought forth a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal.</i>
<i>Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated</i>	<i>Now we are engaged in a great civil war, testing whether this nation, or any nation so conceived and so ded-</i>

Figure 47. Samples from the Ayres Handwriting Scale. By permission of the Publisher Cooperative Test Division of Educational Testing Service, Princeton, New Jersey.

of the pupil's handwriting. In the Ayres scale (see Figure 47 and Appendix B, p. 472) all specimens are from the same paragraph of the Gettysburg Address and pupils must write this paragraph under specified conditions. Most of the existing scales, like the Ayres, have a single series of samples that cover all grade levels and are progressively more legible and attractive.

Although a product scale provides a fairly reliable comparative index of the adequacy of a pupil's handwriting, it does not provide the detailed information that may be of use in instruction. For this, it is necessary to observe a pupil as he writes and to analyze his written products in terms of arm and hand movement, letter formation, spacing, etc.

Speed of handwriting is easily measured by timed observation of pupils as they write various assignments or by timed tests. In the latter procedure it is well to use a familiar passage so that speed of thinking or of reading may not be confused with speed of writing per se. According to Freeman a two-minute period is optimum for such a test (29). It is important that pupils start and stop on time, that the test passage be similar to the material they usually write, and that it contain all letters of the alphabet in about the same proportion as they normally occur.

Evaluative Standards in Handwriting

The standards usually applied to an evaluation of handwriting are the norms of handwriting scales, the letter and word forms cited as ideal in handwriting manuals and, as always, the standards in the minds of teachers. The

first type of standard permits evaluation of a pupil's status relative to his peers and to usual grade expectations. Keeping in mind a pupil's ability and his opportunity to learn, his progress in handwriting could be judged good to poor as he exceeded or fell below the appropriate norm. There is at least one published scale, the *Conrad Manuscript Writing Standards*,¹⁰ which can be used as an evaluative standard for manuscript writing.

The letters and words in manuals and on wall charts used as models by pupils as they learn to write afford a different sort of standard. As examples of what is correct, they may be used to tell a pupil how well he is forming letters and words and where he needs to improve; but they should not be used as a basis for marking. As compared with the models, the handwriting of all third graders is a failure and not one eighth-grade child in a thousand deserves an *A*. Yet both third and eighth graders might be learning to write with great efficiency and thus deserve high marks.

What any teacher thinks constitutes poor and good handwriting at any grade is a vague standard and is easily affected by what the teacher thinks of the pupil as a whole. Since objective standards exist in product scales and in the specimens of manuals and wall charts, dependence on a subjective standard is thought to be unnecessary and unwarranted.

In lower grade reporting, handwriting may be given a separate mark or comment. In upper grades and in secondary English classes it seldom is marked separately, but it affects language art and English marks to an unknown extent. For the most valid evaluation, it should, of course, receive its own specific evaluation.

LITERATURE

So far, we have dealt with the skills aspect of language arts instruction and measurement largely has been simple and direct. Now, in literature, the other phase of language arts, we encounter phenomena whose measurement is often complex and always indirect. Instruction relative to literature begins as soon as pupils can read for pleasure and for vicarious experience. In most schools informal and incidental attention begins in the fourth or fifth grade; by the seventh grade there are required readings and anthologies may be used, and from the eighth or ninth grade on, study of literature may consume half or more of the time allotted to English.

Dimensions of Literary Achievement

It is certain that all English teachers wish for pupils to learn the names of some books, authors, and characters, to know the difference between poems and short stories, and to recognize what a plot is. Hence, the general dimension of knowledge is an obvious one. Invariably, though, in writing and

¹⁰ New York: Bureau of Publications, Teachers' College, Columbia University.

in discussion, English teachers insist that this is not all they wish to teach nor even the most important thing. They talk about literary appreciation, about improved tastes in reading, and about gaining cultural values from reading as another body of English objectives. So it is necessary to posit another general dimension for literature, literary appreciation. This term is used not because it is a clear one but because it is the term most commonly applied to these objectives. The common quality in them seems to be that of feeling so the term "literary attitudes" might be more appropriate.

Particular courses of study in literature involve varied specific dimensions of knowledge and appreciation. When a teacher wishes to measure his pupils in relation to literature he should focus on specific dimensions most appropriate to his instruction. In the following outline are the types of dimensions with which teachers of literature at all grade levels are likely to be concerned. The outline may be used as a general source or simply as an example of the way dimensions may be stated.

Some Dimensions of Literary Knowledge and Appreciation

Knowledge

1. What and how many things are known about literary works: type, author, title, characters, settings, plot, style, inherent philosophy, lesson or moral, relations to other works, publication events, etc.

2. What and how many things are known about literary history: authors, titles, chronology, schools, influences, socioeconomic relationships, development of forms, recurrent themes, the views of critics, etc.

3. What and how many things are known about literary forms and techniques: anecdote, short story, novel, drama, poem, essay, sketch; types of plot, of rhyme, of meter, of point of view, of characterization, of plot development, of narrative and descriptive styles, of figure of speech; literary modes: tragedy, comedy, satire, melodrama, epic, farce, fable; etc.

Appreciation

1. Preferences in reading, as to titles, authors, forms, subject matter, etc.

2. Purposes in reading: for simple diversion, for escape, for vicarious thrills, for emotional release, for ennoblement, for aesthetic experience, etc.

3. Transferred feelings: during reading, extent to which the emotional content of the reading (if any) is felt vicariously by the reader: mirth, love, rage, sorrow, etc.

4. Transferred values: during and after reading, the extent to which the values expressed or inherent in a work are incorporated by the reader: ethical norms (for example, *the Golden Rule*), political views (for example, *democracy is the superior form of government*), social standards (for example, *youth should respect and obey their elders*), philosophic dogma (for example, *man is a puppet of blind physical forces*), etc.

5. Aesthetic discrimination: extent to which the reader derives pleasure from the form, structure, technique, language, etc., of a work as apart from its content and the extent to which he discriminates among different works according to

their "aesthetic quality." For example, a tenth-grade boy is apt to like adventure stories. His *aesthetic discrimination* would govern to some extent his preferences among these titles: *Tarzan and the Apes*, *Tom Sawyer*, *The Broad Highway*, *Beau Geste*, *20,000 Leagues Under the Sea*, *Tom Jones*, *Scaramouche*, *Kidnapped*, and *Captain Midnight Comic Books*.

The dimensions of literary knowledge and appreciation are noteworthy in several respects. They are essentially arbitrary. They nearly all represent covert behaviors or states of mind and feeling. They tend to lap over into one another and have vague and unbounded definitions. Because of these attributes, careful attention must be given to a statement of dimensions before their measurement is undertaken.

Because they are arbitrary, it may not be assumed that any two efforts to measure achievement in literature are related. When a given teacher or a given standardized test essays to measure this phenomenon, what is measured is what the measuring procedures happen to deal with and not some independent, always-the-same entity, "literary knowledge and appreciation." Hence, it is mandatory to say what is meant by literary achievement in the given instance and to claim only to measure that.

Because the dimensions are covert behaviors, they are not measurable as such and must be "translated" into the overt behaviors directly related to them. This means, for example, that pupil actions must be found which show that a pupil knows that "an epic is a long poem or song about a legendary national hero" and show that he has incorporated "a sense of personal integrity and courage" from reading *Invictus*. An action related to the first item might be to write down the essential features of an epic when asked to do so. An action demonstrative of the second item would be more difficult to determine but might be such a thing as a pupil's "referring to *Invictus* as he persists in an unpopular stand with his classmates."

Finally, because dimensions of literary achievement tend to be vague, it is essential that they be defined as precisely as possible before measurement is undertaken. If, for instance, one dimension is to be aesthetic discrimination, the factors of a work that constitute its aesthetic aspect must be stated: unity of plot and action, variety of phrasing, appropriateness and novelty of figures of speech, naturalness of meter, etc.

Forms and Procedures of Measurement in Literature

Description-classification and ranking are the forms of measurement symbols applicable to literary knowledge and appreciation. Scale numbers are precluded because no test or rating "scale" pertinent to literature has yet been devised that has the attributes of a true scale. Since the dimensions of literary knowledge and of literary appreciation are covert behaviors, their measurement must be indirect and all procedures of measurement except product analysis seem to be applicable.

MEASURING LITERARY KNOWLEDGE

Guided response items useful for measuring literary knowledge are illustrated in Figure 48. The several standardized instruments available for this

29. The character in *Ivanhoe* who spent several days in a coffin, supposedly dead, was

- 29-1 Cedric
- 29-2 Isaac of York
- 29-3 Brian de bois Guilbert
- 29-4 Gurth
- 29-5 Athelstane 29 ()

Literary knowledge is amenable to true-false, matching and ordering items as well.

10. Which one of the following lines changes its meter in the middle of the line?

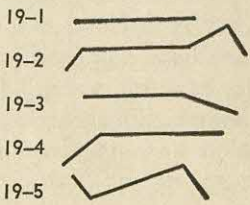
- 10-1 "Take her up tenderly; lift her with care."
- 10-2 "Lovely and airy the view from the hill."
- 10-3 "My love, my love, my love, why have you left me alone?"
- 10-4 "I fled him down the nights and down the days."
- 10-5 "For these red lips, with all their mournful pride."
.....10 ()

A short passage is printed from a work of fiction, an essay, or a poem, and questions are asked such as the following.

17. This passage is notable chiefly for its

- 17-1 beautiful language
- 17-2 colorful descriptions
- 17-3 clever dialogue
- 17-4 continuous action
- 17-5 suspense 17 ()

19. If we were to draw a line to indicate the rise and fall of interest in this passage, it would look approximately like which one of the following?



67. The writer's mood appears to be one principally of

- 67-1 optimism
- 67-2 doubt
- 67-3 anger
- 67-4 sadness
- 67-5 patience

Figure 48. Examples of guided response items for measuring literary knowledge. (All are from Cooperative English Tests for High Schools, published by Educational Testing Service, Princeton. Reprinted by special permission of Educational Testing Service.)

task (see Appendix B, page 470) are useful in ascertaining how much of a given body of information given pupils possess. It is necessary for English teachers to devise their own tests to measure what and how much pupils are learning in a given instructional situation. However, self-devised tests may use the same type of items as the published tests.

In addition to guided response testing, literary knowledge may be assessed through free-response items. They are more efficient for knowledge of works and forms than for knowledge of literary history. The latter may be handled so validly through guided response items that there seems to be no point in using the less reliable and more time-consuming free-response approach. In the case of works and forms, however, such essay questions as "Compare the verse of Sandburg with that of Whitman" and "Distinguish between poetry and doggerel" are likely to provide much better evidence of what the pupils know than a series of multiple-choice or true-false questions.

Perhaps the most efficient single procedure for measuring how well pupils comprehend both the substance and the technique of a particular work is to have them read it under controlled conditions and then to answer either guided or free-response questions relative to it. This device, illustrated in Figure 48, is frequently used in standardized tests with summaries being substituted for lengthy works. If a series of such tests are given which include passages embodying many different literary forms and techniques, pupils' knowledge of forms and techniques in general may be assessed. One of the best tests produced during the Progressive Education Association's evaluation of experimental High School programs in its "8-Year Study" (22) was based entirely on O. Henry's story, *A Municipal Report*.

MEASURING LITERARY APPRECIATION

Measurement of the other basic dimension of literary achievement, appreciation, is a far more complicated process and has been notably unsuccessful. The two published tests that seem to have the greatest promise have no norms and provide no data as to reliability and validity. These also were products of the "8-Year Study's" evaluation staff and now are available

through the Educational Testing Service. One, a *Check List of Novels*, asks students to indicate whether or not they have read a series of novels and whether they liked them or not. From the pattern of any student's answers, his literary habits and tastes presumably will be revealed. The other, *Inventory of Satisfaction Found in Reading Fiction*, requires that students confirm or reject a series of statements about what readers get out of reading or what they dislike. Again, the pattern of a pupil's responses is to be analyzed, this time with a view to determining *why* he reads fiction or *what significance* it has for him.

Literary Preferences. Both of these tests and any other direct approach to measuring literary preferences is in danger of the "school solution" error. In many areas, certain attitudes receive more social acceptance than others. Pupils long have learned what the accepted attitudes are and when asked by teachers for their attitudes they usually assert the right ones. They have learned that this makes teachers happy and that it keeps the pupil out of trouble. So pupils are a little more inclined to *say* they like the classics than they actually do and to say that their free reading motives are aesthetic or moral, whereas actually they may be to seek thrills or to escape.

Observation and free-response procedures are useful in measuring literary appreciation. Reading tastes can be determined by watching pupils in the library when they have time for free reading; by asking pupils to list the books, stories, and articles they read over a period of time; and by asking such questions for free response as, "What is the best book you've read and why? What is the worst book you've read and why?"

Significance of Reading. The significance of reading for a given child or youth is one of the most difficult of all factors to determine. Simply asking him in a test "Why do you like to read?" is certain to produce answers heavily weighted in the direction of how he thinks you want him to respond. If there is time, a carefully directed discussion with a pupil about his reading can add to and verify evidence from a test. Of course, what a pupil reads is in itself indicative of why he reads, since given categories of literature and titles have known psychological appeals.

Transferred Feelings and Values. Two of the most important appreciation dimensions, transfer of feelings and transfer of values, are least amenable to measurement by tests. The test situation is likely to produce responses that reflect stereotyped viewpoints as well as what the pupil actually derives from his reading. Moreover, since feelings and values often lack the unity and uniqueness necessary for clear remembering, a recollection of "how I felt as I read this" is apt to be greatly distorted no matter how conscientious the pupil is.

For appraising these two dimensions it is necessary to rely mostly on observation and self-evaluation. Pupils, particularly the younger ones, suggest their feelings by the degree of their attention and by their facial expressions as they read. Their talk and action after they have read a work are partially

indicative of any values they have incorporated from the work. But, for the most part, only the pupil can know what he feels and what values he has derived from reading. He can be taught that such transfer of feeling and value is a proper outcome of reading and that he can gauge his own appreciation of a work by noting how he responds to it.

Aesthetic Discrimination. Measurement of aesthetic discrimination may be approached by asking pupils to select among passages that vary as to aesthetic quality or by determining whether or not they know what are the aesthetic factors in any work, or in general. The first approach is illustrated by the following item.

Which lines of poetry do you like best?

As passed the noon and came the eve
— a. Then dressed all things as they did grieve.

Now came still evening on and had
— b. All things in her somber livery clad.

As evening approached all things turned
— c. Dark—the servants of the night.

The second means of measuring aesthetic discrimination is typified by this type of item.

Check the factors essential for good poetry

- a. Exact meter
- b. Pleasing rhythm
- c. Alliteration
- d. Aptness of sounds to ideas and moods

- e. Repetition and variety
- f. Noble ideas
- g. Figures of speech

The first approach seems to have the greater inherent validity since the task required is the sort of thing a person does when he exercises aesthetic discrimination. It requires, however, that a prior judgment be made as to the better of several comparable existing lines or passages or that plausible decoys be invented to go with a selected good passage. The advantage of the second method is its adaptability to all aesthetic aspects and the ease with which these can be sampled. On the debit side, it requires an assumption that verbal knowledge of an aesthetic element or principle is evidence of the application of such knowledge to reading and such assumption may not be warranted.

Evaluative Standards for Literature

When a teacher wishes to evaluate pupils' accomplishment in literature, he has little in the way of external standards to help him. Some few of the published tests of literature have grade norms but these have no meaning

beyond the tests themselves. Moreover, the norms from different tests are not mutually equivalent. The nature of literary knowledge and appreciation precludes the existence of any nonrelative standard of correctness or propriety such as dictionaries and style books can provide for composition.

The standards used by a majority of English teachers are for the most part arbitrary and subjective. Letter marks or other symbols of value often are assigned to guided response test scores on the basis of 70 per cent = *D*, 78 per cent = *C*, 85 per cent = *B*, 90 percent = *A*. Here custom and the teacher's arbitrary decision are the source of the percentages—marks relationship. In other cases, a free response test is read and assigned an *A*, *B*, *C*, *D*, or *F* value with no intervening scoring nor any comparison with a known standard. What has occurred has been that the teacher simply and immediately expressed the degree to which he liked or disliked the pupil's answers. Of course, the teacher's liking or disliking is based on his notions of what pupils should know and feel about literature. And in view of the teacher's training and experience, it is a legitimate basis for evaluation. However, the teacher's feelings reflect other things than the pupil's test response and there is no possible way of verifying this type of evaluation or of checking it for error. The problems involved in arbitrary and subjective standards are explained more fully in Chapter 9.

Arbitrary or subjective marking may be minimized by the use of a performance scale as a standard. A teacher, better a group of teachers, can describe the various levels of achievement in literature that pupils are likely to manifest in a grade or in a series of grades. They can have three, five, or however many gradations best suit the differences among pupils. They can give each gradation a letter or number equivalent. As they measure the performance of pupils they can compare their status with the gradations of performance on the standard, find the point of best fit, and assign a value to the pupil's achievement accordingly. This process does not remove subjectivity from evaluation but it does minimize it. Moreover, it is a systematic and rational process, not an arbitrary or emotional one. Such a type of standard and manner of marking are explained in detail in Chapter 9, pages 203–205.

Literature seldom is marked separately on report cards but usually is combined with composition and other things in a single mark for English. It should, of course, be marked separately; and knowledge and appreciation, as the two basic dimensions, should be marked separately. For any detailed and diagnostic evaluation of literary achievement, each separable subdimension of knowledge and appreciation must, of course, be evaluated individually.

SECONDARY SCHOOL ENGLISH

English instruction in junior and senior high schools and in junior colleges involves one or more of the aspects of language arts and, hence, little special

attention needs to be given to measurement and evaluation in English per se. There are two things that must be stressed. First, English is a composite subject and its various parts or dimensions should be measured separately. Second, a number of marks should be issued in English, not just one *A*, *B*, or *D*.

It will do little good, though, to perform separate measurements and evaluations if the pupil and his parents are to receive a report card with a single mark in English. So far as pupils are concerned, evaluations are significant only when they are communicated to them. If a single letter, *A*, *C*, or *F*, is the only symbol used for evaluating such a complex subject as English, the pupil has received an inadequate and possibly misleading communication. Consequently, teachers are urged to use a number of marks, one at least for each major dimension of each language arts element in the course.

In Figure 49 is shown a very simple composite report card for use in

	<i>As compared with ability</i>			<i>As Compared with other pupils in the same grade</i>			<i>As compared with a standard covering all grades</i>												<i>Special comments</i>
	Poor	Acceptable	Good	Below	Average	Above	Grade Equivalent 5 6 7 8 9 10 11 12												
Composition Mechanics Effectiveness																			
Reading Rate Compre- hension																			
Speaking Mechanics Effectiveness																			
Literature Knowledge Appreciation																			

Figure 49. A profile reporting form for English or Language arts.

English. The dimensions of each element are not broken down as extensively as they might be, and as they should be during measurement. However, a further breakdown might confuse the pupil and not justify the additional space and effort involved.

The first rating column equates a pupil's performance with his ability and entries necessarily will be imprecise and subjective. Effort and gain from the beginning of a marking period are judged here. The third column entails use of a standard that has several gradations of competence deemed typical of successive grades. Entries might be made in terms of age rather than grade equivalents. The idea is to evaluate a pupil by telling him that his work is like that of the same, of inferior, or of superior grade pupils. The second column and its use are thought to be self-explanatory. Since so many measurements

will result in rank symbols, these evaluations may require no more than a summarizing or averaging of measurements.¹¹ Special comments should be entered as needed to clarify or supplement the column entries.

If usage in a school requires that a single grade be recorded and issued for English, an *additional composite evaluation* still can be given to the pupils.

Summary

The principal phenomena of language arts so far as measurement is concerned are reading (including reading readiness), speech, composition (including grammar, spelling, and vocabulary), handwriting, and literature.

Reading readiness is measured by observation, ocular examination, and readiness tests directed toward such factors as mental maturity, oral language ability, and sensory acuity. Reading has various dimensions, rate, comprehension, and vocabulary being the major ones, and is measured frequently by standardized tests. Reading grade or reading age is the usual index of a pupil's skill. Many tests yield separate scores for various reading dimensions as well as a total score.

Speech is measured largely by observation, testing being inappropriate for the most part. Of concern to teachers are dimensions of rate, rhythm, pitch, amplitude, tone, enunciation, and diction. Standards for evaluating speech tend to be subjective and based on textbooks and the teacher's experiences. Use of recordings of pupil speech of various gradations of quality is advocated.

As taught in the schools, composition is a complex of actual writing and "separate" subjects of grammar, spelling, and vocabulary. Actual writing is best measured through analysis of pupil writings. Grammar, spelling, and vocabulary may, of course, be measured through this means too, but there are available for these dimensions a number of test procedures and many published tests. Standards in use are dictionaries and style books and teachers' ideas of what a given grade should accomplish.

Handwriting is judged through observation of pupil specimens and by comparison of certain specimens with a series of charts showing gradations of skill and beauty. Because of this direct comparative way of measurement, little attention is given to discrete dimensions of handwriting but among them are letter form, slant, and spacing.

Literature is the most difficult aspect of language arts to measure. The basic dimensions of a pupil's achievement in literature are knowledge and appreciation. Both of these are covert dimensions and must be measured indirectly. Locally devised tests are the mainstay for measuring the knowledge items though many standardized tests are on the market. The attitudes and discriminations involved in literary appreciation may be appraised by attitude scales and by tests of aesthetic knowledge and taste. Evaluative standards are almost entirely a function of the individual teacher.

¹¹ See Chapter 9 for further discussion of report cards and evaluative standards.

Secondary school English comprises most of the aspects of language arts. In evaluating achievement in English, it is important that the separate phases be separately evaluated.

EXERCISES

1. Prepare a card file of published tests for all areas of language arts that concern you. For each test give identifying data (publisher, cost, grade, time, etc.) and a statement about its validity and reliability. Base the latter on an inspection of the test and reviews of it in *Buros' Mental Measurements Yearbook*.

2. Prepare an analysis or rating form to use in evaluating pupil compositions. Indicate clearly the dimensions to be measured, how they are to be measured, and the standard to be used in evaluating them.

3. Prepare an observation form to use in appraising a student's public speaking ability. Include essential dimensions and appropriate ways of rating them. Use the form to rate the speech of one or more students.

4. If you specialize in a subject other than English, list the *special* dimensions of reading and vocabulary important in the subject.

5. For the following abstract dimensions of writing and speaking, indicate tangible things or overt behaviors related to them.

Interest
Originality
Style

Humor
Organization

6. Prepare a guided response test for measuring usage, vocabulary, and spelling.

BIBLIOGRAPHY

1. Buros, Oscar K., *Fourth Mental Measurements Year Book*. Highland Park, N. J.: Gryphon Press, 1953.
2. Caldwell, C. G., "Process of Communication in Children's Letters," *Elementary School Journal*, 49:79-88, October, 1948.
3. Coward, A. F., "Comparison of Two Methods of Grading English Compositions," *Journal of Educational Research*, 46:81-93, October, 1952.
4. Dolch, E. W., "Testing Reading With a Book," *Elementary English*, 28:124-125, March, 1951.
5. Dole, A. A., and Fletcher, F. M., Jr., "Some Principles in the Construction of Incomplete Sentences," *Educational and Psychological Measurement*, 15:101-110, 1955.
6. Ebbitt, W. R., and Diederich, P. B., "Validity of an Examination in Writing," *College English*, 11:285-286, February, 1950.
7. Fels, W. C., "College Board English Composition Test, Present and Future," *Education*, 11:4-10, September, 1950.
8. Free, R. J., "We Measure Growth Together," *Educational Leadership*, 4:464-468, April, 1947.

9. Gerber, J. C., "Testing and Evaluation in the Skills of Communication," *College English*, 9:375-384, April, 1948.
10. Gesell, Arnold, and Ilg, Frances, *Child Development*. New York: Harper Bros., 1949.
11. Gray, W. S., "Reading," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 965-1005.
12. Greene, H. A., "English—Language, Grammar and Composition," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 393-395.
13. Griffin, W. J., and Venable, T. C., "Examination of Examinations," *Education Forum*, 16:211-218, January, 1952.
14. Henig, M. W., "Predictive Value of a Reading Readiness Test and of Teachers' Forecasts," *Elementary School Journal*, 50:41-46, September, 1949.
15. Huddleston, Edith, "Measurement of Writing Ability at the College Entrance Level: Objective vs Subjective Testing Techniques," *Journal of Experimental Education*, 22:165-213, March, 1954.
16. Hurlburt, Delpha, "The Relative Virtue of Recall and Recognition Techniques for Measuring Precise Knowledge of Word Meaning—Nouns, Verbs, Adjectives," *Journal of Educational Research*, 47:561-576, April, 1954.
17. Husbands, K. L., and Shores, J. H., "Measurement of Reading for Problem Solving," *Journal of Educational Research*, 43:453-465, February, 1950.
18. Jordan, A. M., *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapters 5 and 6.
19. McGann, M., "Diagnostic Testing of English," *Journal of Educational Psychology*, 38:499-503, December, 1947.
20. Osburn, W. J., Review of *Hudelson's Typical Composition Ability Scale* (1923), in *Fourth Mental Measurements Year Book*, ed. O. K. Buros. Highland Park, N. J.: Gryphon Press, 1953, pp. 316-317.
21. Preston, R. C., and Botel, M., "Reading Comprehension Tested Under Timed Conditions," *School and Society*, 74:71, August 4, 1951.
22. Smith, E. R., and Tyler, R. W., *Appraising and Recording Student Progress*. New York: Harper Bros., 1942. (Vol. III of *Adventure in American Education*.)
23. Spache, G. D., "Comparison of Certain Oral Reading Tests," *Journal of Educational Research*, 43:441-452, February, 1950.
24. Stanley, J. C., "On the Adequacy of Standardized Tests Administered to Extreme Norm Group," *Peabody Journal of Education*, 29:145-153, November, 1951.
25. Tharp, J. B., Coleman, A., and King, Clara B., "Foreign Languages, Modern," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 479-480.
26. Thomas, Macklin, "Construction Shift Exercises in Objective Form," *Educational and Psychological Measurement*, 16:181-186, 1956.
27. Travers, R. M., "Review of Procedures for the Evaluation of the Outcomes of Teaching English," *Journal of Experimental Education*, 17:325-333, December, 1948.
28. Traxler, A. E., "Measurement in the Field of Reading," *English Journal*, 38:143-149, March, 1949.

29. West, P. V., and Freeman, F. N., "Handwriting," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 524-529.
30. Vordenberg, Wesley, "How Valid Are Objective English Tests," *English Journal*, 41:428-429, September, 1952.
31. Wrightstone, J. Wayne, "New Developments in Evaluating Achievement in the High School," *Education*, 71:210-216, December, 1950.

CHAPTER 11

SOCIAL STUDIES

"Social Studies" instruction throughout the United States consists of a wide variety of courses, subjects, curriculums, and skills. In some places it is a collective noun meaning History, Geography, and Civics. In others it signifies a twelve-year unified study of society and its problems, without regard to traditional areas of knowledge. And, in a great many instances, it represents a unified study of society in the elementary grades and a collection of "subjects" in the secondary grades. Where Social Studies is considered a subject in itself, there tend to be included such "noncontent" items as study skills, critical thinking, and social sensitivity, and even personality adjustment and vocational exploration.

Despite this diversity of elements, there is a National Council for the Social Studies. More and more, high school courses are being designated as Social Studies 1, Social Studies 2, etc., even if they are simply U.S. History, World History, etc. College majors are being offered in Social Studies or the more academic Social Science. There seems to be an increasing tendency among schools and schoolmen to treat Social Studies as a subject *per se* rather than as a means of classification. And, finally, there are many common elements among history, geography and civics when it comes to measurement.

Consequently, measuring and evaluating in the Social Studies is presented here as a unit, but attention is given to achievement in subjects, to progress in units on society, as well as to problem solving and allied skills.¹ Personality and group guidance aspects of Social Studies instruction are excluded since they are more conveniently handled in other chapters, 15 and 16. The structure of the first part of the chapter is like that of the previous one. First, attention is given to the phenomena and dimensions of social studies, then to forms and procedures of measurement, to standards and marking practices, and finally to certain special factors in Social Studies evaluation: the marking of citizenship, the Social Studies unit, and diagnosing disabilities. In the second part of the chapter a detailed description is presented of how measurement and evaluation might actually be conducted in one class.

¹ *Buros' Mental Measurements Yearbooks* (4) use Social Studies as a division heading for standard tests although the bulk of the tests are specific to history, economics, etc.

GENERAL CONSIDERATIONS

Phenomena and Dimensions to be Measured

In our consideration of measurement and evaluation in the language arts we were able to draw heavily on research and on measurement practice for a determination of appropriate measurable phenomena and their dimensions. Here, however, we must rely somewhat more on logical analysis. Published research is rare that deals critically with what to measure in the Social Studies and current standardized tests of History, Geography, etc., permit few clear inferences as to the dimensions they measure.

MAP AND CHART WORK

In nearly all the social studies pupils read and draw maps, interpret and produce charts. Hence, one obvious thing that needs measurement is *skill* in map reading, chart interpretation and similar activities. The more important measurable dimensions of these skills include four of the general ones described in Chapter 2. In maps and charts occur many conventional symbols and arrangements, proper recognition of which is essential for meaning. For example, maps are full of ----- (railroads), + (mountain crests), ✈ (airports), ∞ (passes), ☆ (capital cities), ° (degrees). Charts are read from left to right and top to bottom, the headings of columns and rows apply to the entire stretch of the columns and rows, decimal places are a function of position, and the "legend" explains the meaning of dotted lines, crosshatching, various colors, etc. These conventional symbols and arrangements are the elements of the map and the chart, and two dimensions of the pupil's skill with maps and charts are then:

1. The *identity* of the symbols he knows.
2. The *number* of such known symbols.

In addition, the time it takes pupils to read and interpret or to produce a given map or chart is important and therefore *rate* is a third dimension. Moreover, pupils must read the symbols aright, must follow columns and rows accurately, must judge scale precisely, so *error* is a fourth dimension of the skills. Then we know that certain pupils are map-minded and others are not, certain ones refer to charts in many different contexts while others use a chart only when specifically ordered to do so. Consequently, application or the tendency to use or not to use charts and maps in one or several other activities constitutes an additional important dimension.

KNOWLEDGE OF SUBJECTS

A second phenomenon of Social Studies, even more universal than map and chart work, is also more ambiguous and less tangible. Teachers insist that pupils must learn the facts, develop the necessary concepts, and get the rela-

tionships right if they are to *understand* or to have a good *knowledge* of economics, or history, or local government.² Such knowledge-understanding apparently is the name for a construct, an explanation for certain differences we observe in the behavior of pupils. If John writes a good essay on California climate and Bill a poor one, it is customary to say that John "knows" or "understands" the subject while Bill does not. The "knowledge" is *not*, according to customary usage, the behavior of writing, it is something that "lies back of" the writing, which, quite literally, *understands* the behavior.

Whatever else may be involved, *remembrance* of words heard, read, and spoken, of figures and pictures seen, is critically involved. So for purposes of measurement it may be said that *knowledge* or *understanding* means the totality of a student's remembrance of whatever he has seen, heard, imagined, guessed, or otherwise sensed that relates to the subject in question. A pupil's *knowledge* of the geography of North America would mean, for example, whatever the pupil can recall of the text and reference material, of the teacher's explanations, of his answers to study questions, of motion pictures and slides, of the statements made by other pupils during discussions *plus* the additions, deletions, and distortions his own imagining has devised about the subject.

Given this definition, the measurable dimensions of the phenomena are, first, the identity and number of facts or single ideas remembered and their organization, and the concepts or groups of ideas remembered. A fourth important dimension consists of the feeling correlates of given items of knowledge. Along with remembrances of images come feelings of like and dislike, interest and aversion, confidence and anxiety. Since the significance of a pupil's knowledge-understanding is affected by such feelings, appraisal of them is essential to adequate evaluation.

To illustrate these four dimensions, we use a pupil's understanding of North American geography. Among the components of his knowledge are the recollection that the mean rainfall in the Southwestern deserts is less than 10 inches a year and that Charleston is the capital of West Virginia. Both of these are "facts" or single ideas. Among his concepts are the generalization recalled from a chapter summary, "the seaward sides of mountain ranges tend to be verdant and the landward side arid," and a rule of capital location that he has thought of himself, "state capitals are usually in the middle of states." The pupil's "mind" generally contains a very large number of facts and a somewhat smaller number of concepts relative to these. The organization among his ideas is one of simple verbal classification; he thinks of cities together, then of rivers, then of mountain heights, etc. Finally, to exemplify feeling correlates, the pupil "likes" to recite place names and populations but

² Logicians and even psychologists often distinguish between knowledge and understanding on the grounds that the former has more to do with ideas in isolation while the latter is more concerned with the pattern among the ideas. However, for purposes of measurement the two are hardly separable and we use knowledge to mean both.

feels "uneasy" about what we would call ecological ideas, that people in the great plains have a culture different from that of the residents of Atlantic coast states.

ATTITUDES

A third phenomenon of common concern in the Social Studies is the pupil's *attitudes* toward things of social significance. Measurement of attitudes is examined in detail in Chapter 15. Social Studies teachers primarily are concerned with attitudes toward governmental forms, nationalities, races, peers, school, and other matters important for democratic citizenship.

PROBLEM SOLVING

Finally, *problem solving* is a phenomenon that most Social Studies instructors may wish to measure. In problem solving are included such ill-defined entities as critical thinking, reasoning ability, and scientific method. By problem solving is signified a pupil's behavior in a situation that requires some novelty of action for success. This may be a response to a direction to prepare a notebook, the action a discussion leader undertakes when several of the discussants become aggressive, or simply a pupil's essay on an "original" topic. The dimensions of problem solving are, first, the dimensions of the knowledge, the attitudes, and the skills involved. In addition, to appraise a pupil's status in problem solving with any precision it is necessary to determine his goals, the ways he tries to accomplish his goals, how long he persists in a given effort, and, finally, his errors in goals or means.

We have presented four phenomena commonly subject to measurement and evaluation in the Social Studies. Obviously, other phenomena will concern given teachers in given situations. Moreover, the dimensions indicated for the phenomena are not exhaustive. Others may need to be added to accomplish special purposes of measurement in certain subjects at certain grade levels. As other dimensions are included, very careful attention should be given to the conditions of measurability described in Chapter 2, pages 19-28. You may recall that phenomena are measurable in the degree to which they have dimensions which:

1. Are common to several like phenomena,
2. Provide sensory data,
3. Are clearly defined,
4. Manifest variation,
5. Produce highly similar reactions among many unrelated and impartial observers.

Forms and Procedures of Measurement

In considering now the forms and procedures available for measuring the dimensions just described, we should notice that the phenomena and their

dimensions are primarily verbal. Thus those forms and procedures which measure verbal behavior are likely to be most appropriate and useful in the Social Studies. Moreover, a procedure must be "insulated" against measuring verbal skill if a social studies phenomenon is to be measured and not language. Nowhere else is it seemingly so necessary to keep tests of knowledge from being tests of reading ability.

CLASSIFICATION AND RANKING THE PRINCIPAL FORMS

With our verbal phenomena and dimensions we must rely a great deal on description and classification. Indexes of status in problem solving or map skill often must remain as verbal summaries or at best as gross verbal classifications, average, good, poor, for example. Where it is possible or appropriate to compare pupils with one another we can express their status in terms of rank. "Nancy is in the third quarter. Tom stands last in his class, Bill is at the fifth percentile." But, save in rare circumstances, we cannot use scaling as a form of measurement for Social Studies. Because our phenomena and dimensions are verbal and because our comparisons must ultimately be with other persons, it is extremely difficult to establish a zero or to devise equal increments of difference. Equal increments and zeros or other fixed reference points are, as we have learned, necessities for scaling.

The rare circumstance occurs when a published test of achievement is used that has been designed to yield "standard scores" of some sort. If the test has been very carefully validated, the resulting scores will relate to a fixed mean and score differences at any segment of the scale may be roughly equivalent to the same degree of differences at any other segment.

MEASURING MAP AND CHART SKILLS

The map and chart skills of Social Studies are amenable to measurement through observation and product analysis. You can watch a pupil draw a map or fill in a time line, and thus appraise his concentration span and his speed. You can appraise his map or inspect his chart to determine what map symbols he uses accurately and how much he knows of chart protocol. There are published guided response tests for map skills and several standardized reading tests have sections on map reading.³ Pupils' compositions based on maps and charts can be analyzed for the map and chart reading skill there demonstrated.

TESTS OF KNOWLEDGE

On the other hand, observation and product analysis probably are inadequate for measurement of Social Studies knowledge. To get at the dimensions we described for knowledge and properly to sample its extent, we need the advantages of standard stimulations and, at times, of standard responses

³ See Appendix B, page 489.

in addition to standard analysis systems. Consequently, we must use a great many free and guided response approaches: essay questions, things to list, charts to fill in, multiple option and true-false items, matching questions, etc.

APPRAISING PROBLEM SOLVING ABILITY

Problem solving is most frequently measured through observation. Where situations in which the pupil is observed are *problem situations to the pupil*, this procedure can be effective. Certainly, observation is inherently the most valid procedure for measuring complex behavior. However it is time-consuming and often unreliable.

Problem solving facility can be measured to some extent by properly designed paper-and-pencil instruments. Illustrative of these is the test of "Ability to Apply Social Facts, Generalizations, and Values" contrived and used by the Progressive Education Association in its Eight-Year Study of Secondary School Programs (15). In this test the student is confronted with a problem situation, is required to choose among three courses of action, and then to support his choice with reasons. The student's reasons are analyzed according to a schedule and, needless to say, carry more weight in scoring than his choice of action. A second example of a "test" that deals with problem solving was designed by Edwards to measure development in critical thinking (7). Facts are presented about the "common cold" and the student is required to compare the validity of a number of statements purporting to be based on these facts.

STANDARDIZED TESTS

When you wish to use verbal tests to measure Social Studies phenomena, be they study habits, knowledge, or problem solving, you frequently may choose between constructing your own test and using a published one. The *Fourth Mental Measurements Yearbook* (4) lists forty-eight published instruments for use in the Social Studies, distributed as follows:

Social Studies (general)	7	History	19
Civics and History	1	Political Science	10
Economics	4	Sociology	2
Geography	5		

In addition, most achievement batteries have Social Studies components and many published interest and attitude tests are applicable to certain Social Studies tasks. The titles of certain tests, together with identifying and critical notes are presented in Appendix B, pages 462-492.

In using them, it is well to be reminded that published tests in Social Studies probably were not made just for your purpose nor your class. They have to be generally applicable to be published and therefore may not fit your situation perfectly. Furthermore, the only absolute advantages in a published test are that it may have large population norms and can yield norm scores

while yours will not. But, if you construct your test, use, interpret, and revise it properly, yours can be just as reliable, just as valid, and possibly a great deal more pertinent than the published ones. It usually is necessary to use published tests in measuring many schools or classes and in comparing any given group of pupils with large populations of pupils.

Evaluative Standards

Generally accepted and uniform external standards do not exist in the Social Studies. This is in contrast to the situation in Language Arts with its dictionaries for spelling and pronunciation, its handbooks for usage, and its charts for penmanship. In lieu of generally accepted and uniform standards, typically three sorts of things are used:

1. Statements of instructional objectives
2. Textbooks
3. The teacher's ideas or ideals of how things should be

OBJECTIVES AS STANDARDS

In the case of the first, the objective may be anything from a very general single statement, "Learn about the past and its relation to the present," to a detailed list of statements, "Recognize several primary and secondary sources," "Discriminate among biased and objective sources, etc." The relative adequacy of such course or curriculum objectives as evaluative standards depends on how well they meet the conditions previously established for them (pages 193-194).

To examine the usefulness of curriculum objectives as standards, let us look at the Social Studies objectives of a ninth-grade Basic Studies Course in a medium-sized high school (9), and test these objectives against the characteristics of a valid standard. The objectives are:

1. To acquaint the student with the society in which he lives.
2. To inculcate an appreciation for the social customs of the times and man's cultural heritage.
3. To teach the need for civic responsibility, and to kindle enthusiasm for active participation in our government.
4. To develop the ability to listen actively and to judge critically when ideas are presented.
5. To encourage group activities and co-operative planning.
6. To understand man in relation to his physical environment and to see the need for conservation of natural resources.
7. To "sell" the advantages of the "American Way of Life."

Among the conditions of adequacy established for evaluative standards are reasonable objectivity, clear definition, constancy, expression in terms

comparable to those used in measurement, practicality, and, foremost, they must constitute a clear-cut variation scheme of quality appropriate to differences among pupils.

How well do the objectives meet these criteria for standards? The terms in 2, "appreciation" and "cultural heritage"; in 3, "responsibility"; and in 7, "American Way of Life," have widespread feeling connotations and are involved with the individual's personalized motives and beliefs. All the objectives contain vague terms. Consequently, the objectives are probably too subjective and ill-defined for efficient use as evaluative standards.

It is true that they probably are reasonably constant and are practical. Revision of the course and its objectives should not be anticipated for five years or more. Moreover, the objectives do not seem to represent the caprice of a given teacher or a deviate philosophy. No protocol, time, or expense is involved in their use.

However, the objectives are not expressed in terms comparable to those that will express the measurement of status. Neither do they satisfy the most important condition of all: that they must constitute an appropriate variation scheme of quality. The measures probably will be "75 per cent" on a 30-item test, *B* on a notebook, etc., and such comments about behavior and study habits as "noisy but good-natured" and "doesn't concentrate easily."

The objectives are statements of a given accomplishment level or goal. Any spread or range of individual difference must be implied. For example, ". . . to judge critically when ideas are presented" permits us to say "fine," "OK," "you made it," or some such to students who do whatever we mean by "judge critically." But what should we say about the student who falls short of this level, and of the third student, who we feel is even worse than the second? Are we to say that both are simply "unsatisfactory"?

Now, it is considered that these particular objectives are not atypical of Social Studies course objectives generally. Consequently, it is probable that course objectives as such are not likely to constitute good evaluative standards. They can be the source of standards or a point on a standard scale and, of course, they may be reconstrued to become standards.

TEXTBOOKS AS STANDARDS

Using a textbook as a standard against which pupils are judged is likely to be even more invalid than using course objectives. While they are objective, clearly defined, constant, and practical, they have nothing to do with the differences among pupils or with the quality of pupil achievement. Textbooks are only a source of information or a means of instruction. Their role in evaluation is to constitute one of several compilations of facts and concepts against which the pupil's knowledge may be checked for extensiveness and accuracy. Thus they have more to do with test construction and scoring than with marking.

TEACHER OPINION AS STANDARDS

It is useless to try to validate teachers' ideas against our several conditions for validity. They come in all sizes and shapes and, like small boys, won't stand still. Needless to say, they fail to meet the first condition of objectivity completely and tend to be inconstant. But *they can* measure up to all the others. Whether they do or not is a function of the given teacher. Objectivity and constancy can be approached by writing down whatever concepts of "geographic" or "contemporary affairs" knowledge the teacher intends to use as evaluative standards. To write them down is often sufficient impetus for thinking them through. When this is not done, feelings about pupils (perhaps based on their misbehavior or the teacher's dyspepsia) may be the *de facto* standard.

STRICTLY RELATIVE MARKING

The final current possibility is to use no standard at all. This occurs when test scores are converted directly into grades on some arbitrary or statistical basis. In illustration, the *a priori* decision that 70 per cent is passing eliminates any reference to a standard. The allocation of *A*, *B*, *C*, *D*, and *F* to raw scores on the basis of rank in a class or the characteristics of a frequency distribution is another arbitrary way of judging quality. Here status is merely symbolized in another way. In Chapter 9, pages 194–195, there is further discussion of marking on the basis of a distribution and on arbitrary percentages of test questions answered correctly.

Apparently then, to evaluate properly in Social Studies requires the design of standards and not their selection only. The use of a performance scale containing several defined levels of achievement seems to have the most promise for evaluating knowledge-understanding (see Chapter 9, pages 203–204, for a generalized form of one). In a section to follow (page 279), a performance scale is described for knowledge of U.S. History.

Marking and Reporting

With or without adequate evaluative standards, teachers must mark pupils' achievement in Social Studies subjects and they do so in much the same manner as they do other subjects. Except in the primary grades, the *A*, *B*, *C*, *D*, *F* system prevails for subject achievement. In the primary grades, some sort of "satisfactory," "unsatisfactory" rating together with provision for comments is becoming a fairly common practice. Reporting procedures for Social Studies similarly are little different from those used in other subjects. A single letter mark typically is placed by the subject (fifth-grade Social Studies, eleventh-grade U.S. History, and so forth) and signifies the pupil's over-all achievement in the course for the period reported. In many primary grades, parent-teacher conferences and/or anecdotal statements have replaced the traditional

report card. Teachers comment that such reporting practice is more satisfactory for Social Studies in the primary grades than is letter marking but that it is more time-consuming and requires more skill.

Citizenship and Psychological Grading

All elementary and high schools are concerned that pupils behave themselves properly and that they develop desirable traits of character. Also, the great majority of teachers are basically kind and sympathetic toward their pupils. Salutory as both these conditions are, they greatly complicate the problem of measurement and evaluation. They complicate it in all classes and subjects, but the full brunt of "marking on citizenship" is felt in Social Studies classes.

Practices relative to evaluating citizenship vary from no admitted attention to the matter through lowering of subject marks because of adverse citizenship to a specified grade or grades in citizenship. Where citizenship receives no admitted attention, it is doubtful that it receives *no* attention. The safest assumption in such case is that the teacher's assignment of subject grades is influenced in unknown ways and to an unknown degree by the social behavior of the pupils. Consequently, we have grades with unknown significance: how much is subject and how much is citizenship? Where marks may be lowered for breaches of the peace, a similar situation obtains unless how much the grade is lowered is clearly communicated to the pupil. It is thought that use of a separate grade or grades for citizenship is preferable to either of the other practices and we offer three suggestions as to how it may be done more effectively.⁴

1. Citizenship is a complex. Identify the components and mark them separately: co-operation, honesty, care of property, for example.

2. Observe the pupil's status for these factors, record pertinent events, and base marks on the record, not on recollection or intuition.

3. Let marks represent the pupil's status compared with a standard. This standard needs to be appropriate to the age of the pupils to which it is applied. It should not be the teacher's idealized conception of citizenship for himself or another well-educated adult.

As for "psychological marking," the term is currently used for situations where marks are assigned on the basis of their effect on the ego of the pupil rather than on the basis of his achievement. If a *B* is what the pupil needs to satisfy his desire for success, give it to him. If a *D* will give him the jacking up he needs, assign the *D*. It matters not that the *B* student does inferior work and the *D* does superior work. Marks, in these cases, evidently are intended to be therapeutic rather than informative.

Since the ordinary and usual significance of letter marks is that of value

⁴ The measurement of character and citizenship is examined in greater detail in Chapter 15, pages 415-419.

symbols, you are advised not to engage in psychological grading. To express sympathy for pupils is fine, to motivate them is fine, to have mental health in the classroom is fine, but you can do these things without confusing the evaluation process. Praise the pupil for gain, upbraid him for laziness, and be kind, accepting, permissive in instruction, but let his marks symbolize the value of his achievement according to some known standard.

The Social Studies Unit

The "unit," as usually found in the elementary grades, means a group or series of pupil activities having a single over-all purpose. As a rule, the pupils and the teacher plan their objectives and their activities together; groups of pupils work separately at their own rates to accomplish things connected with the unit objectives; periodically there is interchange of information among the groups; the unit culminates on the completion of group tasks with an all-class activity: discussion, pupil presentations, etc. An example might be a unit with the objective, "Understanding the Indians in Colonial America." One group makes a mural, another an Algonquin Village, a third designs Indian costumes and weapons, etc. The unit culminates with a pageant to which parents are invited.

It may be apparent from this sketch, and it is undoubtedly known by teachers who use the unit, that evaluation here entails a number of special problems and considerations.

First, a unit is a device by which pupils may practice talking, writing, reading, construction, arithmetic, etc., in a "make sense" situation, all in addition to accomplishing the stated objective of the unit. It is suggested that measurement-evaluation of progress in these skills be performed in two ways:

1. Periodically during the semester without any reference to units through appropriate free and guided response testing and product analysis;
2. Casually during the unit through observation.

In the latter case, the teacher is advised to watch extremities of status and *not* to try to appraise all the pupils. During the unit, it is possible to detect disabilities and talents as they may express themselves in the pupils' lives, because of the nonbookish or nondrillish quality of the instruction. And the teacher in a unit situation generally is too busy to evaluate all the pupils on their progress in all skills.

Second, the unit tends to have conceptual and performance objectives or, in the latter case, rules. In the example cited, "knowledge of the food-getting practices of the Eastern seaboard Indians" might have been a conceptual objective, while "to share my ideas with others" and "to be neat in my work" might have been two performance rules. Now these objectives and rules frequently are considered the evaluative standards against which the pupils' learning and actions are to be judged. To use them properly as standards

requires that they be restated according to the conditions requisite for standards. In illustration, "neat in my work" needs to be operationally defined, picks up waste paper, washes out paint brushes, etc., before pupils may validly be judged to exercise a given degree of neatness.

In the third place, much of the pupil work in a unit is diversified as to type and difficulty. In consequence, pupils may not be compared directly with other pupils in the class as a means of determining their status on unit objectives. Moreover, ratings on objectives and rules given to one pupil are not necessarily comparable to the ratings given to another pupil. In the absence of direct comparability, the teacher may resort for classification or ranking purposes to written or remembered stereotypes based on *large numbers* of pupils of a given age who have engaged in similar unit activities.

Fourth, evaluation in the unit often takes the form of group evaluation and self-evaluation rather than teacher evaluation. A justification for these procedures, other than their supposed inherent merits, is that the unit is socialized instruction and that the evaluative aspect of instruction should be similarly socialized. In the view of the authors, the conversations with the teacher about "how I'm doing" generally do help the pupil evaluate himself validly. The other devices, however, the self-rating forms, the group discussion, the committee members' ratings of one another, seem to possess no inherent magic. As employed in some instances they are effective and, as employed in other instances, they do not constitute any degree of measurement or evaluation as we have defined the terms. And by no means should a teacher consider that group or self-evaluation may replace teacher evaluation in unit instruction or, for that matter, in any other mode of instruction.

Finally, we need to recall that the unit has little standardized "content." As a rule, all pupils do not read the same book, use the same references, or study the same topics. Consequently, tests of an informational or factual sort on the subject of the unit are generally inappropriate. To the extent that all pupils do study the same thing, such tests are, of course, effective.

Diagnosis of Disabilities

Diagnosis of disabilities is an infrequent activity in the Social Studies. Publishers' catalogues and test bibliographies list no diagnostic instruments. It is apparent, nonetheless, that progress in history or civics is not along an even front for many students and, hence, diagnosis may be a desirable teaching activity if not a usual one.

Educational diagnosis is a matter of identifying the important factors in a performance and of measuring a pupil's status with respect to them. For example, in Civics, the factors of motivation, personal adjustment, intelligence, reading, study habits, vision and hearing, general experience, as well as his specific instruction and performance under instruction, may be the principal dimensions or variables that determine a pupil's status in this sub-

ject. To diagnose why a pupil is failing in Civics requires measurement of his status with respect to each of these factors and then a careful analysis of the measurements.

Instrumented diagnostic procedures a teacher may use are available for intelligence, reading, and study habits only.⁵ There are certain clinical tests for motivation and adjustment but these require special skill for administration and interpretation. Through observation and questioning some rudimentary appraisal may be made of the pupil's motivation and adjustment and his study habits as well. His cumulative folder and/or a questionnaire addressed to the pupil or his parents may disclose something of his experience. The nurse's records may be consulted for vision and hearing data. An inspection of lesson plans should be informative about what and how the pupil is being taught. Finally, a performance profile in Civics can be drawn through observation of the pupil as he reads Civics material, through appraisal of his compositions and notebooks, through listening to his recitation, and through attention to the types of test items he gets and misses. You are getting at his profile if you are recording phrases like his—"looks at pictures and reads captions carefully, skims the rest"; "mentions specific events and conditions only in composition," "misses 'thought' questions completely." You are not drawing a profile if you are just saying "attends to his reading," "does poorly in his notebook," "fails most tests."

These data (however gross the measures in some cases) enable the teacher to interpret and perhaps properly handle the pupil's Civics situation, not simply judge it. Obviously, making the measurements is laborious and interpreting them is difficult.

EVALUATING ACHIEVEMENT IN U.S. HISTORY IN THE EIGHTH GRADE, A PROTOTYPE STUDY

In previous sections of this chapter the what and how of measurement and evaluation in the Social Studies is discussed in general terms and, in other chapters, general processes and principles of measurement and evaluation are presented in detail. To facilitate application of these generalities to a particular instructional situation, we shall describe now how a teacher might engage in measurement and evaluation in a particular Social Studies class. Since it is to serve an illustrative purpose, the evaluative activities to be described are more extensive than is customary and are somewhat idealized.

Our prototype will be a class at the eighth-grade level because there we are likely to encounter most of the issues of measurement and evaluation that obtain for the Social Studies generally. Let us suppose that the class has thirty-five members, eighteen boys and seventeen girls, and is in a junior high school in a large suburban school district. The school's report card is in terms of

⁵ An annotated list of tests in these and other areas is in Appendix B.

A—excellent, *B*—good, etc., with only one mark permitted per subject but with some space for comments. The class has an age range of 12–6 to 14–8, a range of IQ's from 79 to 133 with a median of 104, and reading levels from the fifth grade to twelfth grade.

Course Objectives. In the first semester of the school year, attention is to be given to colonial development, the Revolution, and the Constitution, or roughly the period 1607–1789. The stated goals of the course of study are to improve:

- a. The pupils' understanding of events and conditions on the North American and to some extent the South American continent from 1607 to 1789, together with their significance for current affairs.
- b. The pupils' attitudes toward government processes, minority groups, and school life.
- c. The pupils' study habits.

Evaluative Dimensions and Standards

According to our protocol for evaluation (see Chapter 9, pages 197–198), the teacher's first step is to establish what is to be measured (the dimensions of achievement) and by what standards they are to be judged. The objectives of the course, since they are stated in behavioral terms, may constitute the three basic dimensions. In addition, each of these has several subdimensions which will be the actual focus of measurement.

1. *Knowledge-understanding of the period 1607–1789.* For this, identity, number, organization, and accuracy of ideas are the primary dimensions. The teacher uses as a standard a performance scale for knowledge-understanding much like the one described earlier, page 204. This consists of five levels of understanding, ranging from that which is unacceptable to that which is exemplary, together with the performances which distinguish each level.

<i>Level</i>	<i>Performance</i>
1. Unsatisfactory	Negligible remembrance of anything read or presented. Inaccuracy is at a high level; 50 per cent or more of pupil's ideas are erroneous.
2. Barely satisfactory	Nearly all statements and answers depend on rote memory of specific things presented and are accurate only for simple ideas: Columbus discovered America, Washington was the first president, etc.
3. Satisfactory	In addition to remembering more facts more accurately, the pupil can make some comparisons, point to some relationships, and make some discriminations among the events of U.S. History. He is inaccurate frequently but usually is aware of it.

4. Good In addition to remembering, comparing, etc., more accurately and more fully than in Level 3, he can explain and interpret with fair accuracy any major series of events or era studied. He can group sequential items properly and express a few appropriate generalizations.
5. Excellent Excels at levels 2, 3, and 4, and shows maximum accuracy for eighth graders. He can do some evaluating; for example, that too many political parties create confusion, and can combine facts and ideas taken from several different sources into a coherent exposition.

2. *The pupils' attitudes.* The teacher notes three dimensions for his second item of achievement, "the pupils' attitudes toward selected current civic and social matters."

1. The number and identity of items toward which the pupils have feelings.
2. The valence (like or dislike, for or against, etc.) of the feeling for each item.
3. The intensity of these feelings.

The question of a standard for attitudes was resolved more simply than was the standard for understanding. On the basis of his experience, the teacher decided what were good attributes for eighth graders. Moreover, he assumed his class was a "normal" one and hence a high position in the class with reference to possession of the good attitudes indicated a high value for the pupil's attitudes and low position, low value. One standard is to be the pupil's status at the beginning of the first semester of the eighth grade and the other is the range of attitude scores for all his pupils at the end of the first semester, all with respect to possession of prescribed attitudes. Value becomes then a function of change from initial status and of rank among peers.

3. *Pupils' study habits.* While seemingly a less imposing factor than the others, evaluating pupils' study habits involves appraisal of a large number of dimensions, study being a complex of behaviors, thoughts, and feelings. The teacher of our prototype class enumerated the following dimensions for study habits:

Reading Skills

1. Identity and number of ideas remembered after reading them.
2. Rate of such reading.
3. Differentials for 1 and 2 according to the following purposes of reading: to remember, to skim, to search.

Study Skills

4. Identity and number of study and nonstudy actions in given period.
5. Components and patterns of searches for materials.

6. Amount and kind of notes taken.
7. Rate of solution to search problems.

Motivation

8. Attitudes on these items: assigned reading, assigned writing, "easy" assignments, "hard" assignments, creative assignments, and new ideas.
9. Attention span: how long (continuous) on a specific idea or task.
10. Study span: how long (discontinuous) on a general idea or task.
11. Identity of self-view as a pupil.

As for standards for study habits, he felt that adequacy here or the lack of it is directly related to achievement and, hence, needs no independent evaluation. He is satisfied to determine the status of the pupils and inform them of it.

Forms and Procedures

As a second step in evaluating the achievement of thirty-five pupils in an eighth-grade Social Studies class, the teacher plans the forms of measurement and the procedures to be used. These are selected on the basis first of appropriateness to the dimensions and standards and then of relative validity, reliability, and efficiency. The procedures adopted are displayed in Table 10 along with the dimensions and standards.

Performing the Measurement and Evaluation

Now, after all this decision and planning, the teacher is ready to do something that actually constitutes measurement. During the first three days of the semester, as he and his new pupils discuss objectives, assignments, class rules, and such, he explains what he intends to do in the way of evaluating their progress. Among the points he makes are these:

- a. You will be graded on what you do in U.S. History, plus what you seem to learn about being a good citizen.
- b. To determine what you are learning and not learning, I will use tests, look at your papers, and observe you. Your job in this class has many parts, so I'll have to judge many things about you.
- c. No one who works as hard as he can should be afraid of failing.
- d. We'll have a period once a week when study will be the order of the day. In these periods each of you will have a chance to talk to me about your work.

INITIAL TESTS OF READING AND ATTITUDES

In the first week, he administers two tests to the pupils. One is a standardized reading test that measures rate and comprehension, has subtests for skimming, searching, etc., and yields scores in terms of reading grade levels.⁶ This is a part of his effort to appraise their study habits. The other test hardly looks

⁶ Published reading tests are listed and discussed in Appendix B, pages 486-489.

TABLE 10

An Illustrative Plan of Measurement and Evaluation
for Eighth-Grade Social Studies

<i>Aspects of achievement</i>	<i>Standards</i>	<i>Dimensions</i>	<i>Forms</i>	<i>Procedures</i>
I. Understanding of given period of U.S. History 1607-1789, etc. (For components, see text, syllabi, and references.)	Five levels of understanding, as described on page 279.	Identity, number, pattern, and accuracy of ideas.	Description-Classification Identify, classify, and count. Identify inaccurate items and express as percentage of total items.	Product analysis (2 papers on topics, 5th and 12th week of semester, notebook covering whole semester, check frequently). Free response (Questions for written answers at end of each unit).
II. Attitudes toward government, minorities, and school life.	Beginning status and range of all pupils relative to attitudes teacher considers desirable.	(1) Identity and number of feeling items. (2) Valence of feelings. (3) Intensity of feelings.	Classification and ranking (1) Identify and count. Rank on basis of summation for each area. (2) Classify + or -. Rank on basis of summation by valences deemed socially desirable. (3) Classify strong, medium, weak. Rank on basis of summation of 2 for strong, 1 for medium, and 0 for weak.	Guided response (List of selected items with provision for indicating degree of like or dislike, administer at beginning and end of semester.)
III. Study habits	No systematic standard.		Classification and ranking	Guided response

TABLE 10 (Continued)

<i>Aspects of achievement</i>	<i>Standards</i>	<i>Dimensions</i>	<i>Forms</i>	<i>Procedures</i>
Reading skills		(1) Accuracy and number of remembrances.	Scores as given by standardized tests.	(Standardized reading test at beginning and end of semester.)
		(2) Rate of reading.		
		(3) 1 and 2 differentials for reading for meaning, skimming, and searching.		
		(4) Identity and number of study and nonstudy actions in given periods.		
		(5) Patterns of search for materials.		
Study skills		(6) Type and amount of reading and discussion notes.	(6) Classify and count.	(4-5) Observation (each pupil twice).
		(7) Valence and intensities of feelings toward given things.	(5) Describe in words.	(6) Product analysis (inspect each pupil's notes twice plus information from notebook analysis).
		(8) Attention span.	(6) Classify and count.	(7) Observation and rating.
		(9) Study span.	(7) As for 2 and 3 for attitudes.	(8) Observation.
		(10) Self-view as pupil.	(8) Time in minutes.	(9) Pupil log.
Motivation			(9) Time in minutes.	(10) Free response by pupils at mid-semester.
			(10) Classify.	

like a test. It is his means of getting the pupil's beginning attitude toward government, minority groups, and school life. Table 11 contains the instrument in abridged form.

When the reading tests are scored, he fills in the profile on the front page and files each pupil's test in his folder. He now knows approximately what to expect from each pupil in the way of reading and is able to identify those who need special help in reading. The attitude tests are scored according to the forms of measurement he has planned and then put in the pupils' files. Since he intends to use the test again at the end of the semester, he neither passes the papers back nor tells the pupils their scores.

FREE RESPONSE TEST ON A HISTORY UNIT

The next effort at measurement is the use of a free response test at the end of the first U.S. History unit, "A New World Comes into the Old." The teacher puts the following questions on the board and gives the pupils forty-five minutes to write their answers.

1. Name the person, persons, or group who originally settled each of the English colonies. Name the colony and after it put those who settled it first.
2. Compare the explorations of the Spanish and the English in the two worlds.
3. Explain the effect of the Crusades on European peoples.
4. Explain why England succeeded in colonizing North America and why the French failed.

The teacher marks the papers by using an analysis schedule of the sort we described in Chapter 5, pages 77-81. The schedule is no more than directions to himself to look for certain things and do certain things to each paper so that he may appraise each pupil's understanding fairly.

- a. Count the separate and relevant ideas in the answer to each question. (You will notice that each question is keyed to a different level of understanding.)
- b. Check for accuracy of statements, write the numbers of inaccuracies for each question, and express as a per cent of the number of relevant statements in each question.
- c. The pupil's total *score* is the sum of the number of relevant statements in each question times the number of the question (representing levels of understanding).

In assigning a letter mark to a paper that has been scored, he looks at the total score, the error percentages, and the score for each question, and compares what these seem to indicate with his standard. He gives a similar test at the end of each unit and these, plus the marks on papers and notebook, largely determine the pupil's grades. A "sample" pupil's paper as he marked it is shown in Table 12.

TABLE 11

Test of Attitudes Toward Government,
Minority Groups, and School Life *

HOW DO YOU FEEL ABOUT THINGS?

	Name		Date			
Directions: Things are listed here that some boys and girls like and others dislike. I want to see how you feel. Show whether you like or dislike each item and how much by putting a check in the proper column. Be just as sincere in your answers as you can.						
	Like Strongly	Like	Indif- ferent	Dislike	Dislike Strongly	
To talk in class						
Football games						
Secret ballots in elections						
To have Negro friends						
Civil Service workers						
Academic subjects						
All members of my crowd to have the same religion						
To get special favors from a teacher						
To write compositions						
Studying						
To vote in elections						
To go to school						
For someone to tell me what to do						
To be with Jews						
To embarrass the teacher						
Politicians						
To make a speech						
Representative government						
Activity subjects						
People who tease strangers						
To compromise						

* This instrument *has not been validated and is not for use*. It is presented as an illustration only.

TABLE 12

A Free-Response Test Scored According To
An Analysis Schedule

L8 Social Studies

Kathy _____

1. Jamestown — John Smith
Massachusetts — Pilgrims
- 5-1 Maryland — Lord Baltimore
Rhode Island — Peter Minuet X
Pennsylvania — The Quakers
- 4 2. Well the Spanish went to South America and Central America and Mexico and the English came to North America. Also the English didn't look for gold as much as the Spanish, and the Spanish were earlier.
- 3 3. The Crusades were battles between the Crusaders and the Moslems over the Holy Land. The Crusaders learned to like spices and other things. I think that they brought back some books and people from Europe learned about the rest of the world.
- 3-1 4. It was because the English won the French and Indian War—because there were more English people—because the English government was more interested in colonizing.

$5 \times 1 = 5$ 20% inaccurate
 $4 \times 2 = 8$
 $3 \times 3 = 9$
 $3 \times 4 = \frac{12}{34}$ 33% "

OBSERVING STUDY HABITS

Shortly after this first unit test, the teacher begins to appraise the pupils' study skills and habits. This is an activity that is to spread over the entire semester since it involves two observations of each pupil. In recording his observations, he uses the schedule shown in Table 13. This instrument serves as a reminder of what dimensions to appraise as well as a record form. One sheet is filled out for each pupil in the first half and again in the last half of the semester. These are filed in pupils' folders and become the basis for class discussions and individual conferences.

Beginning at this time and continuing for the semester, the teacher also rates the pupils' attitudes toward certain study matters. He does the rating on the observation schedule (item 5) at the time of the observation, but bases his judgment on things other than what he sees just then.

In addition to attention span, which he measures by observation, the teacher needs to appraise what he calls "study span." To do this he asks each pupil to keep a log on each of two assignments and turn it in with the assignments. From the logs (see Table 14), he is able to extract an average length

TABLE 13

Observation Schedule for Study Skills and Attitudes

Study Habits			
1.	Pupil's name	Inclusive time of observation	Place Date
2.	List study activities		List nonstudy activities
3.	Describe in sequence the pupil's actions when he searches for a material or piece of information.		
4.	Attention span. Pick a ten-minute period when pupil has started reading or writing and is likely to continue. Time his attentive actions and his nonattentive ones and draw a graph to show how he spends his time.		
Attends to task		Minutes	Nonattentive to task
		1	
		2	
		3	
		4	
		5	
		6	
		7	
		8	
		9	
		10	

Besides the logs, the teacher has the pupils describe themselves as students. About the middle of the semester he asks them to write as much or as little as they think is appropriate in answer to the question: "What kind of pupil am I?" When the papers are turned in, he analyzes them by identifying and counting statements and phrases that clearly indicate each of the following:

Self-confidence	or	Lack of self-confidence
High aspirations	or	Low aspirations
Pride in work	or	No pride in work
Persistence	or	Gives up easily

While this is not a crystal ball, an analysis of a pupil's view of himself as a pupil may disclose factors that are causing failure or maladjustment in school. Where needed, the teacher discusses what he finds with the pupils' as a group or individually.

His final measurement of study habits is made by inspecting the pupil's reading and discussion notes. If a pupil makes no notes, this fact is recorded. If he has kept them, the teacher looks at a three-page sample at the middle and again at the end of the semester. He uses the following self-directions in analyzing them.

- Classify each sentence or separate phrase as a quotation or a paraphrase, as expository or suggestive.
- Count the total in each category.
- Count the words in the three-page sample, omitting prepositions, articles, pronouns, interjections, and conjunctions.
- Ascertain from each pupil the number of pages of reading and/or minutes of discussion the three pages represent.

No total score is derived, but from such an analysis he is able to make a reasonable judgment about the value of note-taking to the pupils' learning. If the judgment is adverse, and it often is for eighth graders, he discusses "what to do" with the pupil. The notes are returned to the pupils but an analysis slip goes into each pupil's file.

MARKING COMPOSITIONS

In the course of the semester, the pupils are required to write two compositions on topics pertinent to the subjects they are studying and these are used as a further means of measuring their understanding of history. The compositions are analyzed in much the same way as are the unit tests.

- Assertions are classified as to the level of understanding they appear to represent and the number at each level is recorded. Each frequency is multiplied by the number of its level.
- Assertions are checked for accuracy and the number of errors is recorded as a percentage of the number of assertions.

The pupil's score is the total of points with the inaccuracy index as a companion score. This score is used, as are the unit test scores, for determination of group statistics. The now analyzed composition is compared with the standard for understanding and the composition is valued accordingly.

MARKING NOTEBOOKS

The third and last approach to evaluating knowledge of history is through an appraisal of the pupils' notebooks. These are compilations of quotes and paraphrases from the text, newspaper clippings, map work, etc. They are distinct from the rest of the pupils' work in that everything in the notebook is at the behest of the individual pupil, not the teacher. The same analysis and evaluation is applied to them as to the assigned papers plus these additions.

- a. Pages are classified as primarily copied or self-devised. Self-devised pages are considered further items at level four or five of the performance scale for understanding.
- b. The different sources or types of source (text, encyclopedia, newspaper, etc.), are listed and the frequency of each source is noted.
- c. Each topic or subdivision is rated as to thoroughness of coverage on a five-point scale.

Measurements *b* and *c* are entered on the reading notes analysis slip as further evidence of a pupil's study skills.

Final Measurement of Reading and Attitudes. The final acts of measurement for the prototype class are a readministration in the last week of the semester of the two tests given in the first week, one for reading and the other for attitudes. These are scored as were their earlier counterparts and compared with the earlier scores. Appropriate notations are then made in the pupils' files. In the case of the attitude instrument, the teacher adds these scores and score differences to the same for previous classes and makes a frequency distribution of the final scores and also of the differences between the first and second scores. Pupils are assigned percentile ranks for status and for change on the basis of these two distributions and their attitudes are evaluated according to the quartile in which they fall.

REPORT CARDS

With measuring and piecemeal evaluating at an end, it remains for the teacher to make out report cards (the summary evaluation). As he works he has constant recourse to charts showing test distributions, to the statements of standards, and particularly to the pupils' files. In general, the periodic evaluations of knowledge-understanding are the primary basis for any pupil's marks; *A*, *B*, *C*, *D*, and *F*, corresponding roughly to the five performance levels of the standard for knowledge-understanding. Evaluations of attitudes serve to raise or lower marks about which he is uncertain. More exact information as to them as well as to study habits is given to pupils separately, in the form of a note or by a conference.

Comments on the Prototype Study

Through a description of measuring and evaluating as a teacher might perform it for a class in Social Studies, we have tried to provide an "operational" experience in measuring and evaluating. You may have noticed that the phenomena and dimensions the teacher measured were not just the same as those we earlier attributed to the Social Studies. Such difference from the average is likely to be true of any Social Studies class and, consequently, measuring and evaluating in any class will be a unique problem. However, our "teacher's" approach and the substance of his procedures are thought to be generally applicable.

While the prototype teacher's performance was intended to be within the scope of any well-educated and energetic teacher, he did engage in more extensive evaluation than most teachers do. Moreover, there is some oversimplification in the example, and by no means are all the *problems* of measurement and evaluation in eighth-grade Social Studies examined. Pupil and parent pressure for grades, "much effort no gain" cases, the "70 per cent is passing" tradition, and the matter of the "normal curve" are among the omitted problems.

Summary

Measurement in Social Studies typically is directed toward pupils' "knowledge" of given subjects (or of the content of units), to their map and chart skills, to attitudes that relate to citizenship, and to their problem solving ability. The measurable dimensions of these phenomena so far as practice is concerned are largely the identity, number, and pattern of elements, e.g., the facts and concepts a pupil remembers, his map vocabulary, and the objects of his strong feelings. Description-classification and ranking are the principal forms of measurement. All measuring procedures are applied to Social Studies instruction with various observation and free-response techniques being the more prevalent.

There are no generally accepted evaluative standards for the Social Studies. In use are statements of objectives in courses of study, the content of textbooks, and the teachers' ideals, all of which have serious shortcomings. In many cases, conscious attention is never given to a standard. Marks in Social Studies subjects are usually the letters *A*, *B*, *C*, *D*, or *F*. These frequently are replaced by descriptive statements and even parent-teacher conferences in the primary grades of many schools. In the Social Studies, perhaps more than elsewhere in the curriculum, citizenship is evaluated and "psychological" grading is practiced.

In the elementary grade Social Studies "unit," evaluation involves a number of special problems. Progress often is reckoned relative to "pupil-planned" objectives. Work is diversified and thus pupils' efforts frequently are not comparable. Pupils as a group may evaluate themselves by discussion. Since there is little standard "content," tests of knowledge usually are inappropriate.

Little attention is given to analyzing disabilities in the Social Studies. However, observation and product or free-response analysis can be valuable sources of diagnostic information.

In the prototype study of measurement and evaluation in an eighth-grade situation, the following were salient features. The teacher planned his program, standards, dimensions, forms, and procedures before he undertook it. During the semester he engaged in some sixteen measurement activities. Many of these involved individual observations of each pupil and analysis of products. The results were collected in a folder for each pupil and were examined by the teacher at the end of the semester as a basis for his mark.

EXERCISES

1. Indicate several measurable dimensions for achievement in one or more of these subjects: U.S. History, World History, Geography, Contemporary Affairs and Problems, American Government.
2. Write a criticism of the evaluative activities employed by the teacher in the prototype study.
3. Prepare a free-response test for understanding of current affairs and problems. Include a scoring plan of the factor rating or counting type for the answers to each question.
4. Devise a guided response test for map reading and interpretation.
5. Construct a vocabulary test for terms important in the Social Studies.
6. Prepare a standard for evaluating understanding of a specific topic in the Social Studies that contains at least four levels of performance. Devise at least one test item appropriate to measuring performance at each level.

BIBLIOGRAPHY

1. Berger, D. W., "When a Class Evaluates," *Educational Leadership*, 6:395-399, March, 1949.
2. Bradfield, J. M., "Whither Secondary School Social Studies," *California Journal of Secondary Education*, 24:179-184, March, 1949.
3. Carr, E. R., Wesley, E. B., and Murra, W. F., "Social Studies" in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 1234-1235.
4. Buros, Oscar K., *Fourth Mental Measurements Yearbook*. Highland Park, New Jersey: Gryphon Press, 1953, pp. 657-673.
5. California State Department of Education, "Evaluating Pupil Progress," *Educational Bulletin*, v. 21, No. 6, 1952.
6. Cummings, H. H., "Trends in the Evaluation of the Social Studies," *Education*, 71:231-236, December, 1950.
7. Edwards, T. B., "Measurement of Some Aspects of Critical Thinking," *Journal of Experimental Education*, 18:263-278, March, 1950.

8. Findley, W. G., "Educational Evaluation: Recent Developments," *Social Education*, 14:206-210, May, 1950.
9. Fleming, Sana B., and Committee, "Course of Study for Basic Studies for Grade 9," Roseville Joint Union High School, Roseville, California, Mimeo., 1954, p. 4.
10. Greene, Harry A., Jorgensen, A. N., and Gerberich, J. R., *Measurement and Evaluation in the Secondary School*. New York: Longmans, Green and Co., 1943, chap. 17.
11. Jordan, A. M., *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1950, chap. 7.
12. National Society for the Study of Education, "The Measurement of Understanding," *Forty-Fifth Yearbook*, Part I. University of Chicago Press, 1946, pp. 1-103.
13. Rath, L. E., "Techniques for Test Construction," *Educational Research Bulletin*, 17:85-114, 1938.
14. Rockowitz, M., "Know Thyself: Techniques in Pupil Self-Evaluation," *High Points*, 33:50-56, October, 1951.
15. Smith, E. R., and Tyler, R. W., *Appraising and Recording Student Progress*. New York: Harper & Bros., 1942. Vol. 3 of *Adventure in American Education*.
16. Thomas, R. Murray, *Judging Student Progress*. New York: Longmans, Green and Co., 1954, chaps. 8-15.
17. Torgerson, Theodore L., and Adams, G. S., *Measurement and Evaluation for the Elementary School*. New York: Dryden Press, 1954, chap. 15.
18. Wesley, Edgar B., *Teaching the Social Studies*. Boston: D. C. Heath, 1942, chaps. 32, 33.

CHAPTER 12

SCIENCE AND MATHEMATICS

If the essential characteristics and the primary contributions of our Western culture were set forth, the activities and ideas embodied in science and mathematics would undoubtedly be high on the list. Our achievements in science and mathematics have helped to form our highly complex and technological society and have enabled us to control our environment for our purposes. Today, we are made even more aware of the essential role of scientists and mathematicians in America as we read about their shortage in industry and government and our failure to train as many of them as other countries. Therefore, it is not surprising to find that the study of science and mathematics occupies an important place in our common schools.

The two subjects are sufficiently different in content and approach to warrant separate treatment. The fundamental difference between mathematics and science lies in their approach. Science is a method of inquiry that is primarily inductive, that is, reasoning from particular events to generalizations or laws. Mathematics, on the other hand, is primarily a method of deductive inquiry, that is, reasoning from general statements or axioms to specific statements.

The relationship between science and mathematics has not always been clear, either historically or educationally. There are those who maintain that mathematics should be subservient to science, and as evidence for their stand they point to the evolution of mathematics from man's social, physical, and religious needs. This point of view has certain educational implications. In schools where it prevails, mathematics would not be taught separately but would be included as a part of other subjects when needed; or at best taught in such courses as business mathematics, shop mathematics, or consumer mathematics.

It is true that mathematics evolved from man's everyday activities and has been and still is an important tool for science. In the nineteenth century, however, the deductive nature of mathematics was rediscovered from the Greeks and since that time mathematics has been making contributions in its own right. A rather spectacular example of the essential difference between mathematics and science, and of the importance of mathematics per se, is provided by the geometry of Euclid. This was developed some 2,200 years ago but is still being studied and used today whereas most of the scientific

theories of Euclid's time are now known to be erroneous. A deductive scheme (mathematics) may stand fairly intact for a considerable length of time while an inductive scheme (science) is constantly undergoing change. The title of Eric Temple Bell's book, *Mathematics, Queen and Servant of Science*, is suggestive of the present outlook on the relationship between mathematics and science. Mathematics provides the rational organization for science and hence is the Queen of Science. At the same time, in the applications of science, mathematics is the tool and hence the Servant of Science.

In this chapter, evaluative procedures in science and mathematics are discussed separately. These separate discussions, however, are combined into one chapter because of the common problems of evaluation involved. These common problems stem from certain similarities of content. Both subjects require special terminology and precision of statement. Both subjects are highly systematized and the ideas involved are closely related and interdependent. Moreover, mathematics and science are mutually involved in many aspects of knowledge and human activity.

For each subject we first shall outline briefly its content throughout the grades and identify some of the more important objectives and dimensions of pupil achievement. Following this, we shall propose some evaluative standards and describe several measuring procedures appropriate both to them and to the dimensions of the subject. Finally, we shall indicate any special problems or considerations involved in the measurement and evaluation of either subject.

SCIENCE

Science has two major aspects. First, it is a body of systematized knowledge that is useful and practical. Secondly, it is a process or approach known as the "scientific method." These phases now are assuming equal importance in the schools, whereas in the past the first had been overstressed at the expense of the second. Certain scientific facts, concepts, and principles should, of course, be a part of basic education so that everyone may live more effectively in his natural environment. At the same time, it is equally essential for children and youth to learn the method or approach science uses in solving problems, acquiring information, and developing laws.

All scientific activity centers around the problem of comprehending and controlling man's environment. Very often the products and end results of scientific activity overshadow everything else concerning science. To many persons, science consists primarily of all the interesting gadgets and machines, new treatments and processes, new materials, and miracle drugs, all of which are given extensive publicity. The real essence of science, though, is found in the activity that produced these spectacular results, and this scientific activity is characterized by the following:

1. Its effort toward accuracy and precision in measurement and description.
2. Discrimination, leading to analysis and classification.
3. Making comparisons and establishing relations that lead to the formulation of underlying principles.
4. Testing hypotheses experimentally and analytically.
5. Developing systems of related ideas using mathematical systems as guides.

So far, we have been discussing science rather than the sciences. We are all aware of the various subdivisions of science taught in school: physics, botany, astronomy, geology, chemistry, and biology. For our purposes, however, we shall concern ourselves only with the two broad classifications of science, namely, the physical sciences and the biological sciences. In the elementary school no special distinction is made among the various sciences and it is not until the secondary level that the branches of science are treated as subjects in themselves.

Science instruction in the modern elementary school centers about the firsthand experiences and the immediate environment of the children. Topics include such items as the changing weather and seasons, clothing, various plants, animals, and other living things, fire, magnets, health, various materials and their uses, and food. In earlier times, the science program in the elementary school was concerned mostly with identifying and classifying things and with observation. Use of the experimental approach and problem solving methods was thought to be too advanced for elementary school children. Today, however, the elementary school science program has been extended to encourage the children to relate and compare their observations and to use the scientific approach in studying materials at their level. We find that grade-school children can carry out simple experiments and proceed systematically to solve problems that are significant to them.

At the secondary school level, ordinarily a biology course or a general science course is required of all pupils. From there on, some specialized courses in science are offered that may be taken as electives or are required for a college preparatory program. Such courses include chemistry, physics, physiology, photography, and agricultural science.

Measurable Dimensions in Science

Formulating the many dimensions involved in science instruction is a formidable task for any new teacher. Fortunately, one can turn to the previous efforts of several committees and groups of persons who have given considerable thought and time to identifying the objectives of science teaching.¹ The

¹ For a more comprehensive summary of individual and committee efforts in identifying the objectives of science instruction, the student is referred to Chapter II of Heiss, Obourn, and Hoffman, "Modern Science Teaching" (7).

objectives stated by deliberative committees are tantamount in most cases to the basic measurable dimensions of pupil achievement. Sometimes they need to be restated in terms of pupil performance. One such group identified the following types of objectives for science teaching (10):

A. *Functional information* or facts about such matters as: our universe, living things, the human body.

B. *Functional concepts* such as: space is vast, the earth is very old, all life has evolved from simpler forms.

C. *Functional understanding of principles* such as: all living things reproduce their kind, energy can be changed from one form to another.

D. *Instrumental skills* such as the ability to:

1. Read science content with understanding and satisfaction.
2. Perform the fundamental (arithmetical) operations with reasonable accuracy.
3. Perform simple manipulatory activities with science equipment.
4. Make accurate measurements, readings, etc.

E. *Problem solving skills* such as ability to sense and define a problem, select a tentative hypothesis after studying facts and clues in the situation, test the hypothesis experimentally, and accept tentatively or reject the hypothesis and test another, draw conclusions.

F. *Attitudes* such as open-mindedness, intellectual honesty, and suspended judgment.

G. *Appreciations* of the contributions of scientists.

H. *Interests* in science as a recreational activity or as a field for a vocation.

The science teacher may use this or a comparable list as a starting point for identifying the dimensions he wishes to measure. Such a list is certainly not to be considered final and should be modified to suit given circumstances. For instance, at the elementary school level the teacher may wish to consider an additional factor, the pupil's awareness and sensitivity to surrounding natural phenomena. Some youngsters are keen observers and seem never to miss anything that goes on around them, while others seem nearly oblivious to their surroundings. This sensitivity and keenness are important aspects in science. Another aspect a teacher may wish to appraise is ability to describe accurately and to formulate precise statements about what is seen.

Any original listing of dimensions needs to be reviewed to determine if any important aspects of achievement are neglected or if there are any dimensions that could be more sharply delineated. If we compare the above list with the list of measurable dimensions common to all education phenomena (see pages 29-30), we find that the dimension of identity and number of components listed in Chapter II is essentially equivalent to the aspects of facts, concepts, and principles presented in the list. However, the dimension, organization of components, mentioned there is not apparent in the list. Conse-

quently, the teacher may wish to add such an item as ability to organize ideas and to see their interrelationship. Then, too, the sixth of the listed objectives, "attitudes such as open-mindedness," etc., certainly must be rephrased if it is to be measured.

It may at times be desirable to devise dimensions from a specific point of view. A good example of this is provided by a group (2) who organized their list of dimensions around the reading of scientific material, this being merely a subheading in the set of dimensions just presented. They set up a list based on any topical reading material in science. Portions of this list are provided below:

1. "*Problems*. Ask a question which requires the student:

(a) to identify the problems to which the statement gives the answer and to recognize the central problem to which a number of statements are addressed;

2. "*Information* (data, laws, principles). Ask a question which requires the student:

(a) to recognize when the information he possesses is inadequate for a given problem;

(b) to indicate kinds of sources of information appropriate for a given problem;

3. "*Hypotheses*. Ask a question which requires the student:

(a) to formulate or recognize hypotheses based on given data or situations;

4. "*Conclusions*. Ask a question which requires the student:

(a) to recognize the generalization(s) involved in an interpretation or conclusion;

(b) to detect the unstated assumptions involved in a conclusion;

5. "*Attitudes*. Ask a question which requires a student:

(a) to recognize in a paragraph or statement proper or improper use of such concepts as causality, teleology, simplicity, consistency, tentative nature of truth, and operationalism."

The dimensions just listed overlap a great deal with the dimensions listed earlier on page 297. However, it is apparent that organizing the objectives of measurement around scientific reading material does place a restriction upon their scope.

Ways of Determining Dimensions. From our discussion so far, there seem to be some four ways for a teacher to develop measurable dimensions, aspects, or objectives in science instruction:

1. Use an already prepared list and make adaptations where necessary for local needs.

2. Identify the items as they currently and persistently arise in science instruction.

3. Compare various proposed lists of dimensions and check for dis-

crepancies. From these lists develop a composite list and check for improvement in delineating the dimensions.

4. If special circumstances require it, take a proposed list or composite list of dimensions and reorganize it from a different point of view.

In devising, selecting, or adapting the aspects of achievement in science which are to be appraised, it is essential to keep in mind the conditions which dimensions must approximate if they are to be measurable. The requirements for measurability are discussed in detail in Chapter 2, pages 19–26. Furthermore, what is to be measured should be known by pupils and hence must be stated so that they can understand it.

Evaluative Standards in Science

Once the dimensions of achievement have been determined, the next step in the evaluation process is to formulate appropriate evaluative standards. As was the case with Social Studies, there seem to be no well defined and generally agreed-upon standards available for use by the science teacher. In many cases the objectives of instruction are themselves the standards. For example, in a chemistry class, the teacher may want all his pupils to learn the names and atomic numbers of the basic chemical elements. To recite all these elements and their atomic numbers correctly constitutes the highest value, to recite less than all of them correctly has less value. In other cases, the mean performance of a class is taken as the basic point of reference for evaluation. Achievement that exceeds the average is marked as meritorious and that which falls short of the average is given a negative rating. Finally, as in many school subjects, evaluation too often is performed without reference to any known standard. The teacher simply feels that the student should have a high grade or a low grade and that's all there is to it.

Our basic viewpoint about evaluative standards is presented in Chapter 9 and, as you may recall, we consider a scale or hierarchy of performance to be the most valid standard for use in evaluating pupil achievement in school subjects. In science instruction, the teacher will need to develop such a performance scale for himself since there are no published ones available. A performance scale is simply a statement of several gradations of behavior with respect to the aspect of achievement in question, which range from that which is of no value to that which is considered to be of highest value. Pupils are judged according to the level their actual performance approximates.

A generalized variation scheme is described on page 204, which may be applied to understanding or knowledge of any subject. Here we shall not attempt to define performance scales for all dimensions of science. This is something that each teacher needs to do himself for the aspects of achievement he wishes to measure. We shall illustrate the process by selecting one dimension unique to science and by developing a performance standard for it.

The dimension we have chosen for the example is "the ability to formulate a scientific experiment that will answer a question or test a hypothesis." This

objective has been selected because it is relevant to both the elementary and the secondary school levels. The illustrations are drawn from the topic of magnets, which is often discussed in the upper elementary grades and, of course, is important in general science and physics at the secondary level. Ordinarily the children read about what magnets can do and then they are given the opportunity to do various things with magnets. To gauge their ability to use an experiment to answer questions about magnets, the teacher raises this question: How can you tell which of these two magnets is stronger than the other? The pupils are then required to formulate an experiment that will answer this question or test the hypothesis that Magnet *A* is stronger than Magnet *B*. In judging their responses the teacher uses the performance scale shown in Table 15.

TABLE 15

Example of an Evaluative Standard for Science

Objective or dimension on which the pupils are to be evaluated: The ability to formulate an experiment that will answer a given question or test a given hypothesis.

Question: How would you find out which of these two magnets is the stronger?

Level I

Performance: Suggests an experiment that misses the point; either it doesn't solve anything or it attacks another question.

Illustration: Pupil suggests that they hammer on the magnets, an experiment which would measure the strength of material in the magnets but not the strength of their magnetism.

Pupil suggests seeing how hard it is to pull the two magnets apart.

Level II

Performance: Suggests an experiment that points toward the answer but it is unrealistic and fantastic in its arrangements, or one that does not control all the variables, so that there still may be some doubt about the answer.

Illustration: Suggests an elaborate arrangement for measuring distance at which magnets attract objects, but fails to account for variation in frictional surface and variation in the material contained in objects.

Level III

Performance: Suggests an experiment that is simple, direct, and realistic, and would give a clear-cut answer to the question or provide a clear-cut basis for accepting or rejecting hypothesis.

Illustration: Suggests having the magnets pick up iron objects of the same shape but of different weights. The magnet picking up the heavier object would be the stronger.

These performance levels cited are designed as evaluative standards for elementary pupils. With refinement, though, they may be pertinent to secondary grades as well.

Forms and Procedures of Measurement in Science

Forms. As we observed in the first chapter, the form in which measures are to be expressed is primarily a function of the dimensions appraised and of the standard to be applied to its evaluation. In science instruction the bulk of the dimensions lend themselves only to classification-description or to ranking. The precise units and fixed reference points necessary for scale measurement are not possible at present except for a few minor aspects of science achievement—rate of performance for one. Since the standards we have described as most appropriate for science evaluation are themselves classification schemes, the use of classification symbols would seem to be most appropriate for measurement. Of course, pupils may be ranked relative to any performance level, or, by interpolation, relative to the whole range of the performance scale. For example, five pupils judged to be performing at the third level of the standard shown in Table 15 could be assigned rank among themselves. If the three levels were considered merely three designated points on a continuum of performance, thirty pupils might be ranked in terms of their position on the continuum.

Applicable Procedures. The many types of procedure for measuring behavioral phenomena—observation, product analysis, free response tests, guided response tests—are described in detail in earlier chapters (III–VI). Our presentation here is limited to a few applications of these procedures to scientific subjects, particularly to those test items and observational methods that can appraise the critical thinking and problem solving aspects of science. (For more comprehensive survey of measurement in science, see the 45th and 46th Yearbook of the National Society for the Study of Education.)

GUIDED RESPONSE ITEMS

So far as knowledge of facts or remembrance of nomenclature is concerned, all types of guided response items are appropriate: true-false, multiple-choice, matching, arrangement, labeling, short answer, etc. All these are illustrated in Figures 9, 10, 11. Two of the most common types of science test items are shown in Figure 50.

An ant is a (an)	The three types of levers are:
a. amphibian	1. _____
b. insect	2. _____
c. mammal	3. _____
d. reptile	

Figure 50. Two usual types of guided response items used in science.

Since these kinds of questions are easily constructed, it is understandable that questions of a factual type often predominate in science tests.

During the past twenty years under the leadership of such men as Tyler, Rath, Noll, and Zechiel progress has been made in developing procedures that show promise of extending the measurement of science beyond the factual dimension to that of scientific thinking. Among the "new" dimensions it is now possible to measure are identification of cause and effect relationships, detection of unstated value judgments, recognition of valid and invalid reasons, determination of whether or not statements can be verified, and perception of limitations of data. In Table 16 we illustrate testing procedures appropriate to these newer aspects of science achievement.

TABLE 16

Examples of Procedures for Measuring Ability to Think Scientifically

Illustration 1. Identifying valid cause and effect relationships.

Below are listed pairs of events. In the blank space between the two events, write:

- A. If the first event is the cause or contributing cause of the second event.
- B. If the first event is the effect or result of the second event.
- C. If there is no cause and effect relationship.

<i>First Event</i>		<i>Second Event</i>
1. Clouds are moving in the sky.	_____	Heavy dust is in the air.
2. A weight is placed on the end of a suspended spring.	_____	The spring stretches.
1. Johnny's eyes water.	_____	Johnny's feet sweat.
4. Walking under a ladder.	_____	Getting hit on the head with a paint bucket.
5. The sound from a drum.	_____	The beating of a drum.

(NOTE: Cause and effect relationships can be very complex and caution should be exercised in setting up this type of device.)

Illustration 2. Detecting unstated assumption necessary to reach certain conclusions.

A certain city reported that during the year 1956 there were 1,216 men drivers involved in auto accidents while only 327 women drivers were involved in car accidents.

Conclusion: Women are safer drivers than men.

What important assumptions are necessary in order to arrive at this conclusion?

(NOTE: This could be a free-response question, or a list of statements could be provided and the pupils asked to check the ones they believe to be necessary assumptions.)

TABLE 16 (Continued)

Illustration 3. Recognizing valid and invalid reasons.

Smith and Tyler and the Evaluation Staff in their book, *Appraising and Recording Student Progress* (15), have made a careful analysis of this dimension. Their procedure for measuring it consists of describing a situation involving natural phenomena, first asking the pupils to give their answer to the problem in the situation and then requiring them to check the valid reasons for their answers from a list of statements. As one example, they describe a situation in which during warm weather persons not having refrigerators often wrap their bottles of milk in wet towels and keep them where there is good circulation of air. The pupils are first asked if this would be effective, and then they are asked to check the reasons for their answer.

In their analysis, they have identified the following categories of wrong reasons that might be present along with the valid reasons for consideration by the pupils.

1. True reason but irrelevant to the problem.
2. Teleological statement indicating some predetermined design.
3. Ridiculous statement.
4. Assuming the conclusion.
5. Unacceptable analogy.
6. Unacceptable authority.
7. Unacceptable common practice.
8. Superstition.

A teacher of science should be able to use this measuring procedure at all levels. A considerable number of commonplace observations can be explained by pupils on the basis of principles learned in their science classes. An example of how the device may be applied to another observation is provided as follows.

Observation: It takes longer to cook food in boiling water at an elevation of 6,000 feet than at sea level.

Directions: Below is a list of statements that attempt to explain the above observation. Check the statements that are valid explanations.

- _____ 1. The relative humidity is less at high altitude than at sea level.
- _____ 2. Food weighs less at high altitude.
- _____ 3. Atmospheric pressure decreases with increase in altitude.
- _____ 4. People should not try to live in high altitude places.
- _____ 5. Temperatures are lower at higher altitudes.
- _____ 6. As the atmospheric pressure decreases, the boiling point of water decreases.
- _____ 7. Homemaking experts recommend that foods should be boiled longer at higher altitudes.

Many other items similar to this example could be developed. The approach could also be used in connection with classroom demonstrations by having pupils check the valid reasons for what had happened in the demonstration.

Illustration 4. Determining if a statement can be verified scientifically.

Directions: For each of the statements listed below, write:

- A. If the statement can be scientifically verified.

TABLE 16 (*Continued*)

- B. If the statement is a theory or hypothesis and hence cannot be scientifically verified.
 C. If the statement contains a value judgment and hence cannot be scientifically verified.
 D. If the statement is a definition and hence is not subject to scientific verification.

- _____ 1. The eye is an organ of a human body.
 _____ 2. The evaporation rate of this pan of water will increase if heated.
 _____ 3. Molecular action in a body of water will increase if heated.
 _____ 4. The earth travels around the sun.
 _____ 5. Mosquitoes should be destroyed because they are disease-bearing.
 _____ 6. Osmosis is a process by which the molecules of a substance pass through cell membranes.

If students can sort out those statements that can be verified and those statements that cannot, and tell why, they are on their way to a good understanding of science.

Illustration 5. Ability to perceive relationships in scientific data and to perceive the limitations of scientific data.

These two abilities have been carefully analyzed by Smith and Tyler and the Evaluation Staff (15) as aspects of a broader ability, namely, the ability to interpret data. They have determined a scheme for classifying students' interpretative statements which are indicative of their perceptiveness for relationships and again of their awareness of the limitations of data.

Relationship interpretations are classified as reading points, comparison of points, cause, effect, value judgment, recognition of trend, comparison of trend, extrapolation, interpolation, sampling, and purpose.

Classifications for the second ability, perception of limitations, are as follows: accurate, overgeneralization (goes beyond data), undergeneralization (overly cautious), and crude error (missed the point entirely). The following exemplifies the measurement of these abilities in terms of these classifications.

Data: A national survey of accidental deaths was made in the U.S. with the following results reported by types of accidents.

<i>Types of Accidents</i>	<i>Percentage of Accidental Deaths</i>
Motor vehicle	36
Railroad	4
Air transport	0.5
Falls	24.5
Burns	6
Drowning	7
Firearms	3
Poisoning	3
Miscellaneous	16

TABLE 16 (Continued)

Directions: Below are some statements that are interpretations of the above data. For each statement, write:

T—If the statement is a reasonable interpretation of the above data.

U—If there is insufficient evidence supporting the statement.

F—If the above data contradict the statement.

(Reading points)

_____ 1. 36 per cent of the people in the United States have cars.

_____ 2. 6 per cent of the fatal accidents involved burns.

(Comparison of reading points)

_____ 3. More people ride the train than the airplane.

_____ 4. Airplanes are safer than trains.

_____ 5. There were more accidental deaths by drowning than by burns.

(Cause)

_____ 6. Accidental deaths by drowning are the result of swimming in unsafe places.

_____ 7. Fatal accidents from falls are the result of carelessness.

(Effect)

_____ 8. If all car drivers were more careful, there would be fewer fatal accidents.

(Value judgment)

_____ 9. Possession of firearms should be severely restricted.

_____ 10. A campaign should be carried out warning people about fatal accidents from falling.

Students' answers are compared with the key answers and each answer is judged as accurate, beyond data, overcautious, or crude error according to the following scheme.

Key Answer

Student's answer	<i>T</i>	<i>U</i>	<i>F</i>
<i>T</i>	Accurate	Beyond data	Crude error
<i>U</i>	Overcautious	Accurate	Overcautious
<i>F</i>	Crude error	Beyond data	Accurate

These comparisons would give an indication of the student's ability to recognize the limitations of data. To measure his perceptiveness of various types of relationship among data, he could be required to indicate the type of relationship each statement involves.

So far, we have confined our discussion to guided response test items. The many possibilities of this type of measuring device have only been suggested here and the interested person is encouraged to investigate the other sources suggested earlier in this chapter. We shall now turn briefly to another category of measuring procedure, observation, and the use of anecdotal records and rating procedures in connection with it.

OBSERVATION

Direct observation of the pupils as they go about their activities in science is a particularly valuable procedure in appraising the status of pupils in science classes. Earlier in the chapter it was mentioned that awareness of and sensitivity to one's physical and biological environment is an important dimension of achievement in science. In informal class discussion it is possible for the teacher to note that some pupils are much more observant than others. This may be made a matter of record by writing short summaries of exactly what happened. A second-grade teacher, for example, might record the following after a class discussion on plants.

When the class was asked to tell how some plants were different from others, Eric was able to list about ten different ways, some of which were rather subtle. For example, he mentioned that some plants had bright shiny leaves, that some plants had soft stems, and that some plants had single stems coming from the ground while others had more than one stem coming up from the ground.

An accumulation of such observations can be the basis for evaluating different children in regard to their sensitivity to things around them. Recorded observations are also useful in appraising such factors as scientific attitude and scientific approach to problems. The teacher would need to record incidents that indicate Mary's open-mindedness or Frank's rigidity in regard to new information. He also should record how Bob set about finding out why a certain crystal radio kit wouldn't work.

Check lists, as mentioned in Chapter 4, page 52, are frequently used in observation to direct the observer's attention to various facets of the thing being measured. Such general dimensions as scientific attitude and the experimental approach to problems have many aspects. Scientific attitude, for instance, comprises such items as a willingness to accept the tentative nature of explanations, an avoidance of becoming "ego-involved" in the experimental process, an unwillingness to distort or to compromise the truth in any way, and an unwillingness to jump to conclusions on the basis of insufficient evidence. These are only a few of the elements of a scientific attitude and their observation is made more efficient if the teacher has a check list of them. In addition, use of a rating scale for each aspect (see page 56) may enable more systematic and comparable appraisal of pupils.

Projects of various sorts are an essential part of science instruction. These include collections of rocks, insects, and plants; the preparation of displays

showing life cycles or working parts of machines; the construction of models, radios, electric motors, and other mechanical and electrical devices; and the reports of field trips.

The teacher might appraise these projects by using a check list and/or rating scale that contains such factors as originality and creativeness, scientific approach, colorful display, and clarity. In laboratory courses, the student's ability to set up the necessary equipment for experiments is an important aspect of his achievement. In appraising this ability, consideration should be given to such aspects as proper choice of equipment, manipulative ability in setting up the apparatus, neatness, and stability, and the over-all effectiveness of the apparatus to do the job it was set up to do.

The illustrations just presented should indicate that direct observation has an important place in measuring and evaluating the competence of pupils in science classes.

STANDARDIZED TESTS IN SCIENCE

The standard critical reference for published standardized tests is the *Mental Measurements Yearbook* edited by Buros (1). The following number of standardized tests in science are listed in the fourth edition of the *Yearbook*:

<i>Subject</i>	<i>Number of Standardized Tests</i>
Elementary Science	3
General Proficiency in Science	4
Biology	11
Chemistry	16
General Science	7
Geology	1
Physics	11

From the above list we can readily see that nearly all the tests are at the secondary or college level and that practically three-fourths of the tests are in chemistry, biology, and physics.

An inspection of some of the more recent tests in science reveals that an effort is being made to extend the coverage of these tests to include other than factual objectives. For instance, in the *Manual of Directions for a biology test* the following statement is made (Appendix B, page 490).

This test has been developed primarily to measure understanding and the ability to apply knowledge in the interpretation of situations and the solution of problems. Testing of ability to recall minute, isolated facts has been minimized. Rather, the student is given an opportunity to demonstrate how well he can discern relationships between what he has learned and the world of living things which he encounters every day.

Tests such as this, which have extended their questions to include the application of principles and interpretations of everyday situations, are consistent with the trend of best practice in science instruction.

Standard tests have much to offer the science teacher, particularly at the secondary level, if certain requirements are met. In the first place, they ordinarily are prepared by experts in the field and much thought and study go into the construction of each item. Secondly, a standardized test has been tried out and revised to meet certain criteria of difficulty and internal reliability. Finally, norms usually are available so that a teacher may compare the standing of his class with other comparable groups of pupils.

Precautions in Use of Standardized Tests. Before these advantages can be realized, however, certain requirements must be met.

1. The objectives and topical coverage of the standardized test should be approximately the same as those of the teacher's class.

2. The students on whom the test was standardized should be comparable to the students in the teacher's class. The norms of a test that was standardized on pupils of a rich urban school district are not likely to be applicable to pupils in an impoverished rural district.

3. The test should be constructed and validated according to prevailing best practice. This may be checked by an inspection of the test's manual and by reading critical reviews relative to it in journals or in the *Mental Measurements Yearbook* (see Chapter 16, pages 426-427), for a discussion of the selection of standardized tests).

Special Problems in Science Measurement

As a final consideration, it is necessary to mention several problems in measurement and evaluation particular to science.

1. Although excellent work has been done in identifying the various aspects of science instruction that should be measured, selection of dimensions for any given class and the development of a standard in terms of levels of performance are left to the teacher in most cases.

2. The science teacher in particular is confronted with the task of evaluating the ability of pupils to carry out a scientific approach to problems and their scientific attitudes. From previous discussion it is apparent that this is a difficult task. In order to do an effective job, the teacher may need to plan some controlled situations that will allow for observation of performances keyed to these dimensions.

3. Ordinarily, the elementary school science program includes a multitude of experiences for the children, science tables, field trips, free reading as well as assigned, and a great deal of incidental relation and discussion. Consequently, the elementary teacher cannot evaluate the pupils' learning of a well-organized body of subject matter. He must isolate the "scientific" aspects of achievement in all this variety of activity and devise means of evaluating them. For this purpose, observation and the maintenance of cumulative records probably are the best recourse.

MATHEMATICS

We come now to a consideration of measurement and evaluation in mathematics. In general, we shall follow the same outline that was used for science. First, the nature of mathematical activity, both in general and in the schools, will be discussed. The broad dimensions of the subject will be pointed out next and some specific breakdowns of these dimensions will be exhibited. Following this, we shall develop sample evaluative standards for several of the dimensions as well as some appropriate measuring procedures. A brief discussion of standardized tests in the field and of some of the special problems involved in the measurement of mathematics instruction will bring this section to a close. In many instances, what has been said earlier in this chapter in connection with science applies here also.

Nature of Mathematical Activity

Mathematics is in nature a highly symbolic subject. Its origins are found in the development of symbols by our primitive ancestors to represent quantities and collections of objects. To deal with quantities, various abstract number systems were developed, culminating in the present-day Hindu-Arabic system, which is especially abstract in that it involves the idea of positional notation. In order to cope with quantitative problems, such operations as addition, multiplication, subtraction, and division were devised. These operations are even more abstract than the symbols and are very closely regulated by a set of rules. Paralleling the development of a symbolic scheme for handling quantitative relationships, arithmetic and algebra, was the development of a symbolic scheme for handling space relationships, geometry. It was in geometry that the Greeks discovered the deductive nature of mathematics. However, the full implication of their discovery was not realized until some two thousand years later. So until present times, mathematics has been considered primarily a useful symbolic tool for dealing with quantitative and spatial problems, and it has been characterized essentially by manipulations of these symbols.

Today there seems to be a strong movement toward a proper emphasis on the long overlooked deductive nature of mathematics and toward a better use of its potentialities. When the content and method of mathematics are carefully examined, it is apparent that mathematics is an idealized language, permitting the greatest possible precision in thought and operation. Intuitive ideas are given symbolic form, and the relations and operations among these ideas are identified and are likewise given symbolic form. Starting with a few assumptions and using certain basic rules of logic, a symbolic system is developed that has myriad ramifications and applications. This system may happen to be arithmetic, algebra, geometry, analytic geometry, calculus, or topology. The method is the same for all.

This increasing emphasis on the deductive nature of mathematics does not

imply, however, that the instructional methods in teaching mathematics should be formal and deductive. On the contrary, the deductive structure of mathematics can be taught most successfully by using informal inductive methods. This has been shown to be true particularly in the "meaning approach" used in arithmetic. Here the students are provided experiences through the manipulation of concrete and semiconcrete materials, and are encouraged to discover for themselves the basic principles, rules, and generalizations that form the deductive scheme of arithmetic.

To conceive of mathematics as an idealized and precise language allows a broader application of its principles and operations. Heretofore, the primary purpose of mathematics instruction has been to teach the symbolic manipulations necessary for the solution of man's physical and social problems. This narrow utilitarian view is now being replaced by a much broader viewpoint which still emphasizes the application of mathematics. Mathematics can be used to facilitate thinking in general. Its symbolic systems are models of systematic, precise, and logical thought. Other subjects and areas of activities use mathematics as a guide for organizing their findings into useful and effective systems.

Mathematics has the same potential use for the individual as well. A person studying mathematics not only can acquire a set of highly useful manipulative techniques but also he can acquire a highly useful model for precise expression of symbolic construction, for identifying relations between ideas, making valid logical inferences, and in general for organizing and using his ideas and experiences. Mathematics can provide a vital experience in how ideas and concepts are interconnected, interrelated, and interdependent, and how the usefulness of concepts depends upon the structure of which they are a part. Through mathematics it can be shown clearly that ideas do not stand alone and that a hodgepodge of ideas is nearly useless. This broader utilitarian view of mathematics instruction is gaining wider acceptance in our schools today.

Measurable Dimensions in the Study of Mathematics

Once again, as in science, it is possible to refer to the work of a number of individuals and committees as a source for the measurable dimensions of mathematics instruction (see bibliography at the end of the chapter). For convenience, we have arbitrarily established four broad categories of dimensions:

1. Computational and manipulative techniques.
2. Concepts and principles involved in quantitative and spatial relationships.
3. Logical structure.
4. Application to mathematical problems.

Within each of these categories are found numerous specific items of pupil achievement relative to mathematics.

COMPUTATIONAL AND MANIPULATIVE TECHNIQUES

In general, this dimension might be described as: "The ability to perform certain computational and manipulative techniques accurately and with facility." The number of distinct computational techniques covered in arithmetic are too numerous to list, as are the varied manipulations of algebra, geometry, trigonometry, etc. Consequently we shall cite only a few examples.

Counting: number building, decomposition of number.

Adding: basic combinations, grouping, carrying.

Subtracting: basic combinations, borrowing, take-away, equal-additions, and complementary methods.

Multiplication: basic combinations, repeated addition, placing of partial products, factors.

Division: repeated subtraction, trial divisors, long and short methods.

Operations with fractions, decimals, and percentages.

Handling of zero.

Removal of parentheses, transposition, cancelation, etc.

Solution of simple formulas in one unknown.

CONCEPTS AND PRINCIPLES INVOLVED IN QUANTITATIVE AND SPATIAL RELATIONSHIPS

The dimensions in this category are concerned with an understanding of the concepts and principles discussed and developed in the study of the various symbolic systems in mathematics. Again the concepts and principles involved in mathematics are far too numerous to present here. We shall confine ourselves to listing some of the dimensions of understanding that pertain to all concepts and principles. These are the abilities:

1. To associate the concept with an example.
2. To provide an illustration of the principle or concept.
3. To develop a definition of the concept.
4. To recognize a misuse of the concept or principle.
5. To identify the principle when used.
6. To discriminate between the concept and other closely allied concepts.
7. To develop origins and justifications for the concepts.
8. To use the principle to explain what has happened in a situation.

LOGICAL STRUCTURE

As we might surmise from our previous discussion of the nature of mathematics, dimensions in this category have only recently been developed and are only now beginning to receive serious attention in the schools. In dealing with this aspect of mathematics, knowledge of the meaning and significance of certain concepts is essential. Among the concepts are: statements, relations, operations, assumptions or postulates, definitions, implications, laws, proof,

and theorems. At least some intuitive notion of their meaning is necessary for any consideration of the structure of ideas and consequently is a basic dimension for this aspect of mathematics.

For additional dimensions of achievement concerned with logical structure, we draw upon the work of the Joint Commission of the Mathematical Association of America and the National Council of Teachers of Mathematics (9). In their view understanding of logical structure involves the following abilities:

1. To formulate clear and concise statements.
2. To identify the principle or assumption which organizes a given set of statements.
3. To distinguish between words that need defining and words that are not to be defined.
4. To identify and formulate the implicit assumptions in an argument.
5. To follow a deductive argument.
6. To organize statements into a coherent logical structure.
7. To formulate a logical argument.
8. To identify and formulate implied relationships between statements.

It is well to note that these dimensions can be applied to nonmathematical as well as to mathematical material.

APPLICATION TO MATHEMATICAL PROBLEMS

In this category come various aspects of problem solving and mathematical applications. Relating mathematics to practical situations is an important phase of the study of mathematics, and for this reason we shall consider its particular dimensions.

In applications of mathematics, measurement plays a vital role. In fact, measurement provides the connecting link between the symbolic systems of mathematics and the physical and social applications. In the problems and applications discussed here, we shall assume that the necessary measurements have already been made and are provided or are readily available. Since measurement plays such an important role in mathematical applications, some understanding of the particular units of measurement that are involved will be necessary. Such units are pounds, quarts, miles per hour, horse power, Fahrenheit degrees, cubic feet, etc. The meaning of these units must be understood before certain problems can be solved.

Some of the dimensions identified for the application of mathematics are the following abilities:

1. To select the necessary facts for the solution of problems.
 - a. By telling what additional information is needed.
 - b. By telling what information may be discarded.
2. To formulate a problem from a given set of data.

3. To estimate the answer to a problem.
4. To symbolize the components of the problem.
5. To recognize the mathematical processes required for the solution and to set up the step-by-step operations.
6. To perceive and express symbolically the unknown and the conditions for finding the unknown.
7. To translate physical quantities and relations to mathematical symbols and relations.
8. To develop a plan for solution (a search model).
 - a. By considering analogous problems.
 - b. By providing the necessary definitions involved.
 - c. By varying the conditions of the problem.
 - d. By identifying patterns and sketching figures.
9. To generalize the problem and generalize the solution.

Evaluative Standards in Mathematics

Texts on mathematics education and instructors' manuals have been far more concerned with what students should learn and how they may best be taught than with how their progress may be validly evaluated. The bases for grades on the report card, for weekly marks, and for the teacher's satisfaction in or concern over any pupil's progress are much the same as for other subjects. Arbitrary percentages of problems answered correctly probably still constitute the most common basis for marks; 95 per cent and better equals *A* and so on to where 65 or 70 per cent is the barely passing mark. In many classes the use of percentages of correct answers has been abandoned in favor of an established relationship between rank in class and a given value symbol. With such a system, the median or mean achievement of the class usually is the fulcrum for evaluation, higher ranks receiving the more favorable marks, and lower, the less favorable. In other situations, quality of performance is reckoned in terms of mastery of a given set of operations, propositions, or theorems. Finally, the norms of standardized tests or of departmental tests sometimes are employed as the basis for judging the value of pupils' work.

As we have observed in discussions of standards for other subjects, the use of evaluative standards such as these leaves several questions unanswered about the pupils' competence. We still do not know what any pupil actually can do who gets an *A* or a *C*. Nor do we know precisely what the *F* student cannot do that earns him the *F*. The best we can say is that the *A* student is probably better at mathematics than the *C* student, and the *C* student better than the *F* student. And if we are thoroughly familiar with the content of mathematics in the grade or course in question, we can surmise in general terms that the student should know and be able to do such and such. But the evaluative symbol based on percentage, rank mastery, or test norm does not in itself indicate this.

For evaluations of progress in mathematics that are informative about the

quality of the pupils' achievement, it is necessary to use some type of performance scale representing gradations or levels of performance from that which is considered to be of least value to that which is of greatest value. A generalized form of such a standard was presented for understanding of any subject in Chapter 9. Earlier in this chapter one was described for a dimension of science and in Chapter 11 a variation scheme was developed for evaluating understanding of U.S. History. Now we will describe one for use with each of the categories of mathematics dimensions just presented. Examples of performance at various levels will be in terms of relevant mathematical problems and/or their solution. Since, as we shall see, the primary procedure of measurement in mathematics involves the administration and scoring of selected problems, the following illustrations of standards afford illustrations of appropriate measuring procedures as well.

COMPUTATIONAL TECHNIQUES

In the category of computational techniques, performance levels are developed mainly on the basis of the seriousness of the errors made in computation. Accuracy and facility are also factors to be considered. The following example is based on a very specific computational technique, that of subtracting numbers in which there are zeros in the minuend (the top number). This is commonly taught in the third or fourth grade.

TABLE 17

Example of an Evaluative Standard for Arithmetic Computation

Dimension: The ability to perform subtraction in the case where there are zeros in the minuend (top number).		
Level	Performance	Examples
I	No apparent idea of what to do about the zero in the minuend. May make some attempts that are completely wrong.	$\begin{array}{r} 406 \\ -235 \\ \hline \end{array}$ (Subtracts 0 from 3) $\begin{array}{r} 231 \\ \hline \end{array}$
II	Can subtract correctly when no borrowing is necessary or when at most only one-step borrowing is necessary. Cannot subtract correctly when two-step borrowing is necessary.	Can subtract correctly: $\begin{array}{r} 406 \quad 340 \quad 409 \\ -203 \quad -215 \quad -352 \\ \hline \end{array}$ Cannot subtract: $\begin{array}{r} \text{(Fails to borrow second step correctly.)} \quad 405 \quad 600 \\ -258 \quad -368 \\ \hline 157 \quad 242 \end{array}$
III	Can handle with facility and accuracy the zeros in various positions, both in the minuend and in the subtrahend. Can also handle two- and three-step borrowing situations correctly.	Can subtract correctly: $\begin{array}{r} 302 \quad 600 \quad 700 \\ -275 \quad -389 \quad -508 \\ \hline \end{array}$

The performance levels have been described in terms of the seriousness of the errors involved. This basic error variation is considered to be keyed to varying degrees of understanding of the computational technique. Accuracy and facility are not specifically included. However, accuracy may be accounted for by variation in careless errors, and facility may be gauged by the variation in time required to complete the exercises. Only three performance levels have been identified here. After some experience with this scheme, the teacher undoubtedly will wish to expand and refine these levels and perhaps add additional ones.

In many cases a teacher, in measuring ability to subtract such numbers, might simply construct a number of exercises involving zeros in the minuend and then evaluate his pupils on the basis of the total number of exercises that were answered correctly. Such a procedure is open to question. For one thing, it could be that the test provided for no basic error variation and hence did not measure variation in understanding of the subtraction technique. Starting with an evaluative standard composed of performance levels, however, should insure that a test will be devised to measure variation in performance.

CONCEPTS AND PRINCIPLES

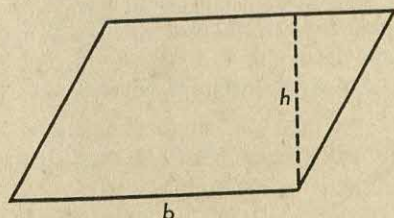
Since the dimensions in this category are connected primarily with the *understanding* of the concepts and principles involved in quantitative and spatial relationships, we may apply the generalized standard for understanding presented in Chapter 9. To illustrate performances typical of the several different levels of understanding, we have selected the concept of the area of a parallelogram, a subject commonly taught in the seventh and eighth grades.

TABLE 18

Example of an Evaluative Standard for Understanding a Mathematical Concept

Level	Performance
I	Can repeat the formal definition for finding the area of a parallelogram provided in class. Can also duplicate the figure drawn previously to illustrate finding the area.

Example:



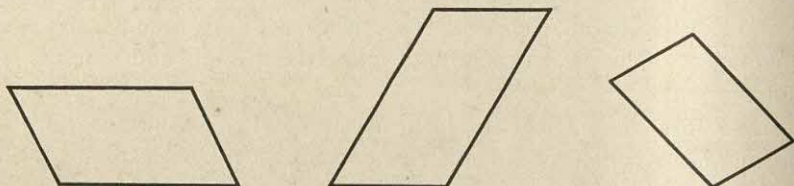
"The area of a parallelogram is equal to the base times the height."

$$A = b \times h$$

TABLE 18 (Continued)

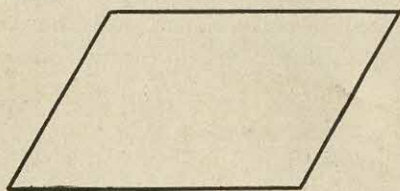
- II Can identify the process of finding the area of a parallelogram in different circumstances.

Example: Show how you would find the area for each of the following parallelograms.



- III Can compare and relate the process of finding the area of a parallelogram to finding the area of other plane figures.

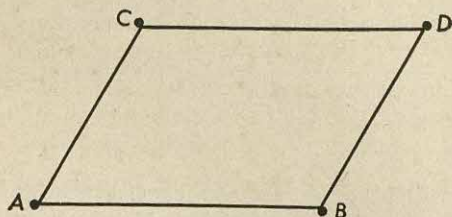
Example:



1. Draw a rectangle that has the same area as this parallelogram.
2. Draw a triangle that has the same area as this parallelogram.
3. Draw a triangle that has half the area of this parallelogram.

- IV Can predict and explain on the basis of his understanding of the area of a parallelogram.

Example:



Points A , B , C , D are all hinged so that the parallelogram can be changed in shape but points A and B are fixed in position.

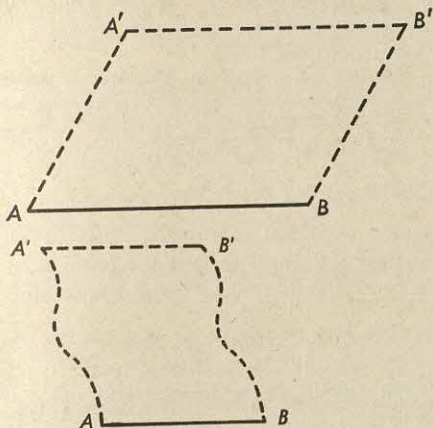
1. Suppose that points C and D are moved to the right with points A and B fixed in position. What happens to the area of the parallelogram? Why?
2. Suppose that points C and D are moved to the left, with A and B fixed in position. What happens to the area of the parallelogram? Why?
3. Draw another position for C and D where the area of the parallelogram is the same as above. Give reason.
4. Suppose that the height of a given parallelogram is doubled. What happens to the area of the parallelogram?

- V Creative. Can reorganize the concept of the area of a parallelogram on a different basis.

Example: *

* This concept of area is discussed by Fehr in *Secondary Mathematics* (3), pp. 6-17.

TABLE 18 (Continued)



Start with line AB in position shown. Then move it to position $A'B'$, keeping the line horizontal and fixed in length. What is the area of the space that was covered by line AB ?

See if you can continue with this idea and develop the areas of other figures. For instance, what is the area of this figure?

The levels of performance described in Table 18 are assumed to be cumulative. In other words, Level III presupposes performance at Levels I and II. This example should indicate some of the many possibilities of the variation scheme for evaluating understanding presented in Chapter 9.

LOGICAL STRUCTURE

To illustrate the use of performance levels as evaluative standards for the dimensions of logical structure we have selected the dimension: ability to detect flaws in reasoning and to identify a valid argument. This dimension is significant at both the elementary and the secondary level. Moreover, the content for the examples in Table 19, the topic "rate of travel or speed," is as frequently encountered in the elementary school as in the high school.

Again it is assumed that the performance levels in the standard are cumulative. It should be noted that the gradations of performance represent the detection of progressively more subtle flaws in reasoning. The levels and examples are meant to be illustrative rather than definitive. To develop a standard for use with a class, a teacher should list many different types of fallacious reasoning, try them out in several tests, observe how many students and what students detected each one, and then devise a performance scale to include the most sharply discriminating flaws.

APPLICATION OF MATHEMATICS

This so-called practical dimension of mathematical ability is a particularly difficult one to evaluate. It is inextricably interwoven with the problem solving process in general, it may involve terms and concepts from many fields of knowledge, and it is often dependent upon ability to read well. If a pupil solves an application problem is it a function of his ability to apply mathematics or is it perhaps a function of his intelligence, his general knowledge, and his skill at reading as well as his ability to apply mathematics?

TABLE 19

Example of an Evaluative Standard for Logic

Dimension: The ability to detect flaws in reasoning and to identify a valid argument.	
Level	Performance
I	Fails to identify any flaws in an argument, even the most obvious flaws. Does not investigate the structure of an argument but judges the validity of an argument only on the basis of whether the conclusion is true or false in his estimation.
	Example: The following argument is presented, "It takes longer to go to the post office than to go to the bank from school. Therefore the post office is further from the school than the bank is."
	Pupil believes this to be a valid argument because he assumes that the distance from the school to the post office and from the school to the bank is the only factor affecting speed of travel.
II	Can pick out the more obvious flaws such as identifying the missing step in an argument and recognizing the irrelevant use or misuse of a definition or principle.
	Example: Pupil recognizes that the argument presented in Level I is invalid by identifying the statement missing, namely that the rate of travel must be constant and equal in both cases.
III	Can pick out the more subtle flaws of an argument such as: assuming that the converse or opposite of an implication is true, using a <i>non sequitur</i> , and begging the question.
	Example: The following argument is presented: "If the car was traveling 60 miles per hour, then the driver was exceeding the speed limit. But the car was not traveling 60 miles per hour, therefore we can conclude that the driver was not exceeding the speed limit."
	Pupil recognizes that this is an invalid argument because of the fallacy of assuming that the opposite of an implication is true.

In keeping with this general difficulty in evaluation, the development of a performance scale for use as a evaluative standard is especially difficult. No generalized one seems possible since each application of mathematics is a function not only of mathematics but of the application situation as well. Particularized ones could be developed for each specific application but their number then would be legion. For these reasons, we shall not provide an illustration of performance levels for this dimension. Instead, we shall indicate

certain types of variable that may serve to determine the efficiency with which a student applies his mathematical knowledge to other areas.

Among the possible variables are:

1. Generalizations, principles, laws known.
2. Extensiveness of manipulative skills.
3. How abstract are conceptions of mathematical operations?
4. How many elements can be manipulated in reaching solutions?
5. Skill at symbolization.
6. Discrimination between relevant and irrelevant operations in any problem.
7. Facility to reverse a process learned in one way.

Forms and Procedures of Measurement in Mathematics

As with science, classification and rank symbols are the most widely used forms for expressing measures of achievement in mathematics. They are appropriate both to the nature of the dimensions and to the evaluative standards. Scale numbers have somewhat more of a place in mathematical measurement than in science. So far as the manipulative dimensions are concerned, instruments may be devised to yield scores that may be treated like scale intervals. More perhaps than in any other school subject is it possible to construct test items that have equivalent difficulty. For example, if the items are all addition problems involving three two-place figures, it may be assumed that any problem is as difficult as any other. Hence, the difference between a score of 70 and a score of 80 may be assumed to be equivalent to the difference between a score of 30 and a score of 40, and the scores may be added, means determined, etc., with mathematical validity.

All procedures of behavioral measurement are applicable to mathematical dimensions. However, because of the exact nature of the material, guided response tests are particularly appropriate. As you may recall, the gist of a guided response item that measures achievement (true-false, multiple-choice, etc.) is that responses mean one thing and one thing only. Nowhere is this condition more likely to obtain than in mathematics, where by definition there usually is only one right answer. Rivaling guided response tests as an effective means of measurement in arithmetic, algebra, etc., is the analysis of pupil assignments done for instructional purposes.

A number of illustrative test items are presented in Tables 20, 21, and 22 for each main branch of mathematics taught in the common schools. They include items appropriate to each of the four basic dimensions of mathematics: manipulative techniques, concepts and principles, logical structure, and applications.

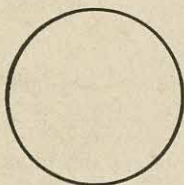
Arithmetic. Table 20 presents a few examples of guided response items in arithmetic selected to show some variety in content and form.

TABLE 20

Examples of Test Items for Arithmetic

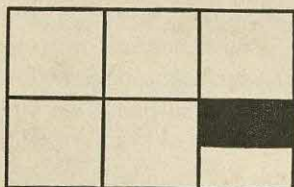
1. In which of the following numbers is the tens digit the smallest? the greatest?
(a) 4,302 (b) 63 (c) 8,563,432 (d) 352
2. If a digit is moved 2 places to the left, how many times has its value been increased?
(a) 200 (b) 10 (c) 5 (d) 100
3. The Cub Scouts sold candy at a carnival. In order to find their profit, list the things you would need to know.

4.



Shade three eighths of this circle.

5.



What part of this rectangle is shaded?

6. In which of the following divisions would the first trial divisor be 5?
(a) $16 \overline{)1,035}$, (b) $51 \overline{)1,035}$, (c) $19 \overline{)1,035}$, (d) $25 \overline{)1,035}$
7. Susan made punch by mixing one cup of orange juice to three cups of ginger ale. What percentage of the punch is orange juice?
Suppose she added another cup of ginger ale. What percentage of the punch is orange juice?
8. If six is added to both the numerator and denominator of the fraction $\frac{3}{5}$, what happens to the size of the fraction?
9. _____ Estimate a line that is $\frac{5}{16}$ of the given line.
10. The population of Circleville is now 35,000. It has increased 20 per cent since 5 years ago. What was its population 5 years ago? If its growth is steady, what will its population be 10 years from now?

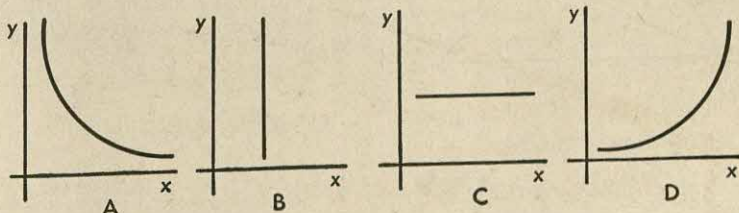
The items shown in Table 20 are the type usually encountered in arithmetic. Further variety can be gained by:

1. Asking questions on portions of a computational process.
2. Asking in words for the computation to be performed rather than having the procedure already set up in symbols.
3. Asking questions about the results of a computation.
4. Changing values of parts of a completed computation and asking what happens to the results.

Algebra. The items shown in Table 21 for algebra are taken from elementary algebra and are concerned primarily with basic definitions.

TABLE 21

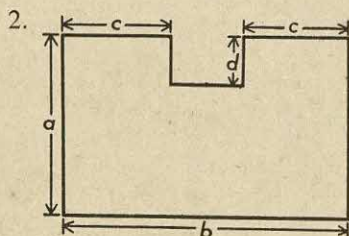
Examples of Test Items for Algebra

1. $k - k =$ (1) $-2k$ (2) $2k$ (3) 1 (4) k (5) 0
 2. $\frac{4x+y}{4} =$ (1) $x+y$ (2) $4x+4y$ (3) $\frac{x+y}{4}$ (4) $x+\frac{y}{4}$ (5) $x+4y$
 3. $s \div s =$ (1) s (2) 0 (3) $2s$ (4) 1 (5) $\frac{1}{s^2}$
 4. $\frac{r+s}{r+s+t} =$ (1) $\frac{1}{t}$ (2) $\frac{1}{1+t}$ (3) t (4) $\frac{2}{2+t}$ (5) None of these.
 5. $\frac{m}{n} - \frac{p}{q} =$ (1) $\frac{m-p}{n-q}$ (2) $\frac{m-p}{nq}$ (3) $\frac{mq-np}{-nq}$ (4) $\frac{mq-np}{n-q}$
(5) None of these.
 6. $\frac{7}{d} - \frac{c}{d} =$ (1) $\frac{7-c}{d^2}$ (2) $7d-cd$ (3) $\frac{7c}{d^2}$ (4) $\frac{7-c}{2d}$
(5) None of these.
 7. $c^3 - c^2 =$ (1) c (2) 1 (3) c^5 (4) 0 (5) None of these.
 8. If m is 6% of n and $m = 12$, then n is equal to
(1) 2 (2) 200 (3) 500 (4) 5 (5) None of these.
 9. What is the value of $\sqrt{d^2 + d^2}$
(1) d^2 (2) d (3) $d\sqrt{2}$ (4) $2d$ (5) $\sqrt{2d}$
 10. A number z is increased by 35%. What is the result?
(1) $0.35z$ (2) $z+35$ (3) $1.35z$ (4) $z+0.75$ (5) $z+\frac{35}{z}$
 11. If $m = \sqrt{\frac{p}{q}}$ then q is equal to (1) $\frac{m^2}{p}$ (2) $\frac{p}{m^2}$ (3) $\sqrt{\frac{p}{m^2}}$ (4) $\sqrt{\frac{m}{p}}$
(5) None of these.
 12. 
- (a) Which of the above curves indicates that as y increases x decreases?
 - (b) Which indicates that x varies directly as y ?
 - (c) Which indicates that as y increases, x remains constant?
 - (d) Which indicates that x and y are everywhere equal?

The questions listed are particularly useful for diagnosing basic errors since most of the common errors are listed as alternate choices. Other items can be developed by using algebraic symbolism to express functional quantitative and spatial relationships.

For example:

1. A is z miles behind B . A 's speed is x miles per hour while B 's speed is y miles per hour. If A 's speed is greater than B 's, how long will it take for A to catch up with B ?

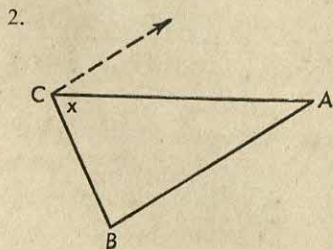
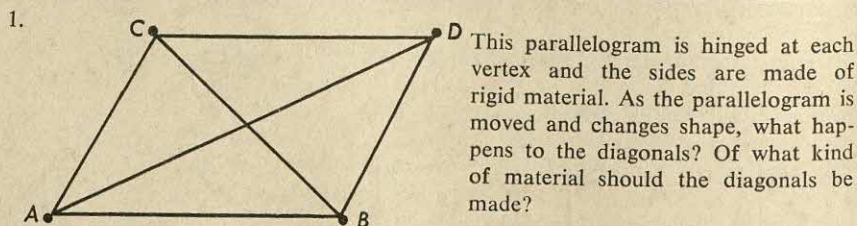


What is the area of this figure?

Geometry. Our illustrations for geometry all involve movable geometric figures that call upon the student to apply his understanding of the properties of these figures, not just to repeat an operation he has learned in class. These questions could be used with students who have been studying intuitive geometry as well as those who have been studying demonstrative geometry.

TABLE 22

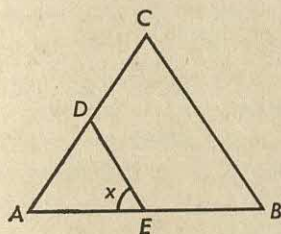
Examples of Test Items for Geometry



The vertex C of triangle ABC moves in the direction indicated parallel to AB , while points A and B remain fixed in position. What happens to the area of the triangle? What happens to angle x ?

TABLE 22 (Continued)

3.

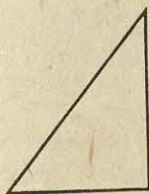


Point E moves along AB while point D moves along AC so that line DE remains parallel to line BC .

(a) What happens to angle x as DE moves toward BC ?

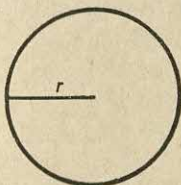
(b) What happens to angle x as DE moves away from BC ?

4.



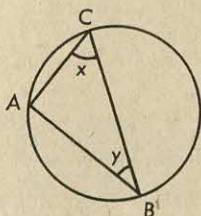
If the three sides of the right triangle shown are each doubled in length, what happens to the size of the angles in the triangle? What happens to the area?

5.



If the radius of the circle shown is decreased 50 per cent, what happens to circumference? To the area?

6.

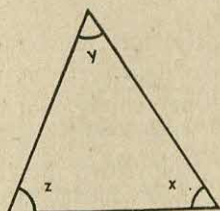


In this figure, points A and B remain fixed in position while point C moves around the circle in a clockwise direction toward B .

(a) What happens to angle x ?

(b) What happens to angle y ?

7.



If angle x and angle y are decreased one half, then angle z is

(1) reduced one half?

(2) increased one half?

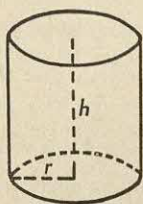
(3) doubled?

(4) the same?

(5) Answer not given?

TABLE 22 (Continued)

8.

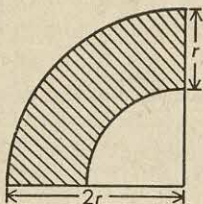


This figure represents a right circular cylinder having a base of radius r and altitude h .

(a) What happens to the volume of the cylinder if the circumference of the base is doubled while the altitude remains constant?

(b) What happens to the volume if the radius is increased 50 per cent and the altitude is increased 50 per cent?

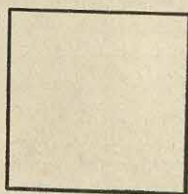
9.



(a) What is the area of the shaded portion?

(b) Suppose r is doubled in length. What happens to the shaded area?

10.



If the perimeter of this square is doubled, what happens to the area?

Many other relational questions can be devised using moving geometric figures. This type of question is particularly effective in testing the students' understanding of geometric concepts in dynamic situations rather than in the static situations usually presented in high school textbooks.

STANDARDIZED TESTS IN MATHEMATICS

The standardized tests available in mathematics are listed in publishers' catalogues and the standard reference work for published tests, *The Mental Measurements Yearbook* edited by Buros (1). The *Yearbook* contains critical reviews of many of the tests and reviews of recently published ones may be found in mathematics and educational journals. Selected tests are cited in Appendix B of this text.

Many mathematics tests cut across subject matter lines. Several cover all phases of high school mathematics and are designed to serve as college entrance examinations. Others are devoted to measuring achievement in general mathematics, including algebra and geometry. Some tests attempt to measure functional competence or the ability to do quantitative thinking. In general, published tests in mathematics usually are employed on a school-

wide basis for purposes of ability grouping, guidance, etc., and by colleges as part of their entrance examination batteries.

Tests of arithmetic measure computational ability primarily, but some include items devoted to arithmetical reasoning. Reviews of more recently published tests indicate that test designers are beginning to emphasize the meanings involved in the manipulative processes and are developing novel approaches to measuring pupils' understanding of them. Some diagnostic tests are available that attempt to help the teacher identify the basic difficulties pupils have in arithmetic.

Standardized tests in algebra include readiness or prognosis tests as well as achievement tests. The prognostic instruments are based primarily on the assumption that a good fundamental understanding of arithmetic is the best predictor of success in algebra. The achievement tests generally give satisfactory coverage to factual content and basic skills. The tests published for geometry tend to be factually oriented and to give too little attention to applications, to inductive development, and to the structure of reasoning and proof.

The following is a summary of the number of published standardized tests in different mathematics areas listed in *The Fourth Mental Measurements Yearbook*:

<i>Subject</i>	<i>Number of Tests</i>
Mathematics in general	15
Algebra	18
Arithmetic	24
Geometry	16
Trigonometry	3

Special Problems of Measurement in Mathematics

Reviewing our discussion of measurement and evaluation in mathematics, the teacher of mathematics seems to be faced with at least three special problems.

1. There is need to avoid the strictly manipulative type of questions where the numbers are all set up for the process to be carried out by the student. It is tempting to compose tests of this type of item since the questions appear ready-made in the textbooks. However, such items usually will measure only rote memory aspects of arithmetic achievement.

2. It is especially difficult to use the closely knit symbolic systems of mathematics in developing questions that show the principles of valid reasoning and the structure of relationship of ideas.

3. It will be necessary for the teacher to devise his own variation schemes or performance scales for evaluating dimensions of mathematics, particularly those involving an application of mathematics to everyday practical situations and an understanding of the concepts commonly taught in mathematics. Pupil

texts and instructors' manuals provide little assistance in this wise, and classroom practice offers few examples of the use of such evaluative standards.

Summary

Science has two major aspects, its body of systematized knowledge and its basic process or approach known as the "scientific method." Many professional groups have defined the dimensions of pupil achievement in science. One particularly useful list includes these items: functional information, concepts, and understandings concerning the physical universe; instrumental skills such as to read science content, perform arithmetical operations, manipulate science equipment and make accurate measurements; problem solving skills such as to define a problem, test a hypothesis, etc., and attitudes of open-mindedness, intellectual honesty, and suspended judgment. Before attempting to measure these or other dimensions, they should be checked against the five conditions established for measurable dimensions in Chapter 2.

In common use as evaluative standards are the objectives of instruction themselves, range and mean of class achievement, and the teacher's general feelings about the value of any pupil's performance. Preferable to these is a performance scale or hierarchy containing several levels of performance, from that which is unacceptable or barely acceptable to that which is exemplary for a given grade of students. Such a standard for evaluating elementary pupils' ability to formulate an experiment that will answer a given question might consist of three levels: plans an inappropriate experiment; suggests an experiment that points toward the answer but is unrealistic in its arrangements or fails to control all the essential variables; proposes an experiment that should give a clear-cut answer to the question.

As in other areas, forms of measurement should be used that are appropriate to the dimensions and the evaluative standards to be applied to them. In science this makes classification and rank symbols the primary forms of measurement. All types of behavioral measuring procedures are applicable to science courses. Guided response items have been devised to appraise dimensions other than the facts-known one. Observation is the most valid procedure for appraising the problem solving ability of students, their sensitivity to their environment, and their proficiency in the laboratory. Many well-designed standardized tests are available for use by teachers but they should be used in keeping with certain requirements. Test dimensions and coverage should be the same as the dimensions and coverage in the course in question. Students in the class should be like those on whom the test was standardized and technical aspects of the test should reflect prevailing best practice.

Historically, mathematics has been taught as a body of manipulative operations and concepts. Today, however, its significance as an idealized language is being realized and mathematics instruction is more broadly conceived than heretofore. Its four basic dimensions are computational and manipulative

techniques, concepts and principles involved in quantitative and spatial relationships, logical structure, and application to mathematical problems.

The commonly used standards for evaluating achievement in mathematics are arbitrary percentages of problems worked correctly, the mean of any class's performance, and mastery of a given set of propositions or operations. Use of a performance scale as a standard is considered superior to any of these. For manipulative techniques, a scale based on types of errors is appropriate; and for concepts and principles, one involving the degree to which a given concept may be adapted to more subtle problems has promise.

All forms of measurement, including scaling for certain tests of computational skill, are applicable to measurement. The principal procedures of measurement employed are guided response tests and analyses of pupil assignments. Illustrations are given for types of items appropriate to arithmetic, algebra, and geometry. Many standardized tests are published for measuring mathematics achievement. Their usefulness is limited by their generalized coverage and their emphasis on computational and operational skills. There is a tendency for more recently published tests to deal with meaning and the logic of mathematics as well as with manipulations.

EXERCISES

1. Develop an evaluative scheme for evaluating knowledge of important scientific facts. What basic dimensions would you have to consider?
2. Examine three standardized tests in science or in mathematics and make a critical analysis of them in terms of what dimensions are being measured.
3. Select a specific dimension which would be most likely to occur in your science or mathematics teaching. Develop an evaluative scheme in terms of levels of performance and then construct a few questions appropriate for each of the levels.
4. In what ways are the measurement and evaluation problems in science and in mathematics instruction alike and different?
5. Why have the dimensions of computational skills in arithmetic been so carefully and systematically analyzed?
6. Select a word-problem in a mathematics text and develop an evaluative scheme you would use to evaluate the answers of pupils to the problem.
7. Develop an observation form that you believe would measure pupils' scientific attitudes in your class.

BIBLIOGRAPHY

1. Buross, Oscar (ed.), *The Fourth Mental Measurements Yearbook*. Highland Park, New Jersey: The Gryphon Press, 1953.
2. Dressel, P. L., and Mayhew, L. B., *Science Reasoning and Understanding*. Dubuque, Iowa: William C. Brown Co., 1954.

3. Fehr, H. F., *Secondary Mathematics, A Functional Approach for Teachers*. Boston: D. C. Heath and Co., 1951.
4. Frutchet, F. P., "Application of Principles," Bulletin 859, PEA, Evaluation Staff of the Commission on the Relation of School and College, 6010 Dorchester Ave., Chicago.
5. Hartung, M. L., and Fawcett, H. P., "The Measurement of Understanding in Secondary-School Mathematics," *Forty-fifth Yearbook of National Society for the Study of Education, Part I*, University of Chicago Press, Chicago, 1946, Chap. 8.
6. Heil, Louis M., *et al.*, "The Measurement of Understanding in Science," The Measurement of Understanding, *Forty-fifth Yearbook of the National Society for the Study of Education, Part I*, University of Chicago Press, Chicago, 1946, Chap. 6.
7. Heiss, E. D., Obourn, E. S., and Hoffman, C. W., *Modern Science Teaching*. New York: The Macmillan Co., 1950.
8. National Council of Teachers of Mathematics, *Twenty-second Yearbook: Emerging Practices in Mathematics Education*. Washington, D.C.: 1954.
9. National Council of Teachers of Mathematics, *Fifteenth Yearbook: The Place of Mathematics in Secondary Education*. New York: Bureau of Publications, Teachers College, Columbia University, 1940.
10. National Society for the Study of Education, *Forty-Sixth Yearbook: Part I: Science Education in American Schools*. Chicago: The University of Chicago Press, 1947.
11. Noll, V. H., "Measuring the Scientific Attitude," *Journal of Abnormal and Social Psychology*, XXX (July-September, 1935), pp. 145-154.
12. Polya, G., *How to Solve It*. Princeton: Princeton University Press, 1945.
13. ———, *Induction and Analogy in Mathematics*. Vol. I of Mathematics and Plausible Reasoning. Princeton: Princeton University Press, 1954.
14. Raths, L. E., "Techniques for Test Construction," *Educational Research Bulletin* (Ohio State University), XVII, April 13, 1938, pp. 85-114.
15. Smith, E. R., and Tyler, R. W., *Appraising and Recording Student Progress*. New York: Harper & Bros., 1942. Vol. 3 of *Adventure in American Education*.

CHAPTER 13

PERFORMANCE-ACTIVITY AREAS

If we were to walk through a school today, we would witness a great variety of activity. In the play yards, athletic fields, and gymnasiums we would watch boys and girls playing games and learning athletic skills. In the shops we would see students busy with the projects in woodworking, auto mechanics, electricity, and sheet metal. In art and music rooms pupils would be painting, drawing, modeling, singing, and playing instruments. Girls in home economics classes would be learning how to prepare meals, to sew, and to care for a family. And so on. From our trip we would have to conclude that a sizable portion of the school's curriculum is devoted to the development of various manipulative and motor skills and that the schools today are "activity conscious."

This chapter is devoted to the general problems of measuring and evaluating achievement in subjects which emphasize motor performance and manipulative activities. These subjects are art, business education, driver education, home economics, industrial arts, music, and physical education. Moreover, we will restrict our treatment to the performances and activities themselves and to the products of these performances. The appraisal of knowledge and attitudinal phases of achievement in these subjects may be performed as it is in any of the so-called "academic" subjects (see Chapters 5, 6, 10, 11, and 12). We shall discuss in turn the measurable dimensions of performances and products in the subjects, the measuring procedures to be used, and standards for their evaluation. The subjects of art, music, etc., will not be treated separately, but sufficient examples will be presented for each to show their unique problems and techniques.

Process and Product. At the outset we need to differentiate between two important aspects of a performance or activity, these being the process and the product. The process in carrying out a performance refers simply to the movements involved and their sequence. The product of the performance, on the other hand, is the result or outcome of the process. The product may be tangible, as in the case of woodworking, or it may be intangible, as in the case of musical selection played on an instrument. Whenever the product of a performance is intangible, the process and product are so intimately interwoven that it is difficult, if not impracticable, to separate the two. Such is the case in singing, swimming, dancing, kicking a football, throwing a baseball, etc.

The degree of emphasis on process or product depends upon the subject and the objectives of the course. In the matter of painting, generally the product receives more attention than the physical movements involved: i.e., holding the brush, mixing the paint, etc. As well, there are courses in which the process or procedure receives nearly all the emphasis. In a beginning golfing or archery class the emphasis may be solely upon form or procedure. In most cases, however, process and product seem to have equal significance. As a rule, the product of a performance is affected greatly by the process that preceded it. Furthermore, the process is determined and modified largely by the desired product in mind.

Measurable Dimensions of a Performance in Progress

We shall consider now some of the general dimensions of any performance which we can reasonably expect to measure. The discussion in Chapter 2 on the essential characteristics of measurable dimensions is pertinent to the problem and it might be well to review them (pages 19-26). In brief, a measurable dimension is relevant to a class of things, manifests variation, provides sensory data, is clearly defined, and produces consensus of reaction among impartial observers.

1. *Speed.* One of the more obvious aspects of a performance is its speed. This is very simply measured by the time it takes a pupil to complete the performance. Comparisons may be made with other timed performances in order to get some idea of how fast relatively this one is. Speed is an essential factor in most competitive performances and in such others as typing and shorthand.

2. *Accuracy.* Another important and common dimension of a performance is accuracy. As a matter of fact, we hear the "speed and accuracy" of performance mentioned more than any other aspects. Accuracy is commonly measured in terms of error counts. This presupposes, of course, that there exists a rather detailed concept of how the performance should be carried out ideally, and any deviation from this ideal is to be considered an error. Several types of errors can be identified, and consequently accuracy is a very broad dimension. Precise measurement usually requires that accuracy be broken down into some of its subdimensions, two of which are listed below.

a. *Procedural errors.* This type of error requires the existence of only one correct sequence of steps or only one pattern to follow in carrying out the performance. It would mean that the procedure has been formalized or standardized by society in general or by special groups. Usually the given procedure has been established on the basis of experience and general acceptance, but in any event a deviation from the procedure would be counted as an error. This would apply particularly to performances governed by generally accepted rules of etiquette, as in serving meals or in dancing, or by legal procedures, as in driving an automobile.

b. *Errors in following instructions.* These errors would occur when the performance is made in response to instructions. Deviations from the given

directions would indicate a degree of noncompliance with the requirements of the task. Errors of this type may be due to actual inability to carry out the instructions or they may be due to inadvertent mistakes. Typing errors are a good example of the latter.

3. *Discrimination*. This dimension involves the selection or choice of tools, equipment, and movements used in carrying out the performance and the perception of stimuli that accompany the performance. Measurement of the dimension is done in terms of adequacy and effectiveness for the operation performed. Discrimination is an important dimension in woodworking, electricity, auto mechanics, and in other crafts where several tools and pieces of equipment are used and where the thing being made or repaired must be perceived carefully and accurately. It is also important in music, where sounds must be heard aright and where tones and finger movements must be selected with exactness.

4. *Economy of effort*. Another important dimension in performance is the economy of effort involved. Here we look for the amount of effective motion as against the amount of "lost motion" or trial-and-error behavior. This aspect of performance is closely related to speed in that the more economy of effort there is, the greater will be the speed of the performance. There is also an element of discrimination involved whenever a choice of movement is made. In most instances, however, the dimension of economy of effort is a matter of well-trained and co-ordinated muscular movement that makes the performance seem easy and effortless. This attribute is particularly desirable in dancing. It is always an essential element in any activity that requires a great deal of muscular co-ordination, such as swimming, playing football, or basketball and gymnastics.

5. *Timing*. This dimension has to do with the rate and emphasis of movement in a complex motor performance. In the operation of a piece of machinery such as a lathe, a crane, or a bulldozer, the several levels and wheels must be operated at exactly the right time and to the right extent. The dimension of "timing" is also involved in any team play, in gymnastics, and, of course, in music and dancing.

6. *Intensity*. Another component of a performance to be considered is the intensity of the action. The outward manifestations of this dimension would be the force or amplitude of the movements involved. Intensity is of particular importance in performances involving strokes, such as tennis, golf, handball, or baseball. The desirable degree of intensity is dependent upon the requirements of the task at hand. It is possible that there may be too much force exhibited or too little, and measurement will have to be in terms of deviation in either direction from some optimum degree of intensity.

7. *Coherency*. This dimension applies to performances in which there is no single correct procedure or sequence of steps for carrying out the tasks involved. In such a case it would be impossible to measure each performance step in terms of adherence to an ideal procedure. Consequently, actions must be judged on the basis of their internal consistency or their mutual appro-

priateness. Measurement of coherency, then, is a matter of the degree to which one step logically follows from a previous step or logically leads to the succeeding step. This dimension is important for woodworking and other shop-type activities involving many individual steps, for which there is no single correct sequence.

Measurable Dimensions of Products of Performances

We need now to consider briefly the aspects of products that lend themselves to measurement. In art are produced drawings, paintings, sculptures, and mobiles. In shop, students make shelves, ash trays, and mail boxes. Dresses and foods are produced in homemaking, and in business education there are type scripts and shorthand notes. Such products have basic physical dimensions (weight, size, color, etc.), which may be measured by the conventional techniques of physical measurement. Of greater interest to the teacher of a performance subject, however, are the dimensions of these products that have to do with their significance or purpose. Among these dimensions are the "composition" of a painting, the "flavor" of a cake, and the "evenness of stroke" in typing. Often the dimensions of most importance to a teacher are the details that have been specified for a product by a pattern, recipe, or set of instructions.

There is far less commonality among the dimensions of products than among the dimensions of performances. Consequently, no effort will be made to define any general dimensions for them as we did for performances in various subjects. The only dimension which approaches general significance is that of accuracy, and even this may not be relevant to certain art products.

Attempts often are made to measure the aesthetic attributes of certain products. This is particularly true in art and music, and often true of products in home economics. The measurement of aesthetic properties is much more complex and difficult than the measurement of physical properties. This is due, of course, to the fact that aesthetic properties often fail to satisfy the conditions essential for measurability (see pages 20-24). If it is necessary to measure any aesthetic dimensions of a product, the measurer must define these dimensions in such a way that they are measurable. For example, "composition" may be an aesthetic dimension of a painting. Before "composition" may be measured other than subjectively, it has to be defined in terms of line convergence, focal point, balance, etc. Such definition of aesthetic properties in measurable terms is illustrated later in the chapter in connection with the design of measuring procedures (see page 345).

Dimensions of Performances and Products in Specific Subjects

So far our discussion of the dimensions of a performance and its products has been somewhat general. We need now to view some of the *specific* dimensions that teachers try to evaluate in subjects where activity predominates. To attempt an exhaustive listing of all the specific dimensions of performance in

each course would be a vain undertaking since their number is legion. Consequently, we shall present only some of the more common ones to be encountered in each subject. These should provide a starting point for measurement in the subjects and, in addition, they should suggest to the teacher other dimensions in which he will be interested.

The sources of the dimensions to be cited are textbooks for the subjects, teachers' guides and methods texts, courses of study, and the few standardized tests published for performance-activity subjects. The dimensions are listed in outline form for each subject. Some are actually subdimensions of others and you may notice that many are merely specific instances or applications of the general dimensions we have been discussing. Furthermore, some of the dimensions, as stated, do not meet the basic conditions of measurability and will have to be redefined in behavioral terms.

Art. First we shall list the specific dimensions commonly mentioned in connection with the *products* of art. The appropriateness of any one will naturally depend upon the medium used.

Art Products

- Organization or composition
- Balance
- Rhythm
- Color
- Contrast
- Repetition
- Sympathy for subjects
- Line quality
- Tone quality
- Spatial relations
- Accuracy of proportion or suitability of distortion
- Stability of subjects
- Ease of interpretation
- Suitability of medium for purpose
- Relationship of proportions
- Textural interest
- Kinesthetic interest
- Technical facility
- Expressiveness and originality

Artistic Behavior

- Perceptiveness to art in everyday living
- Ability to enjoy art spontaneously
- Communication of ideas clearly in artistic expressions
- Ability to criticize and to profit by criticism of art expressions
- Ability to organize artistic forms for certain purposes
- Degree of independence and originality in art expressions
- Application of art principles to activities outside of art class

Music. Several of the dimensions cited for art are equally pertinent to music. These are rhythm, sympathy (for selection), tone quality, ease of interpretation, technical facility, and all those given for "artistic behavior." In addition, there are a number of other important dimensions.

First, much research has been done in studying discrimination in music. Two standardized tests in particular attempt to measure this dimension, namely, the *Seashore Measures of Musical Talent* and the *Kwalwasser-Dykema Music Tests*.¹ In the Seashore test, the student is required to make discriminations for each of the following subdimensions:

Pitch	Timbre
Loudness	Rhythm
Time	Tonal memory

The Kwalwasser-Dykema test measures the following dimensions, which include many types of discrimination:

Tonal memory	Rhythm discrimination
Quality discrimination	Pitch discrimination
Intensity discrimination	Melodic taste
Feeling for tonal movement	Pitch imagery
Time discrimination	Rhythm imagery

Subdimensions of accuracy are important in music and the dimensions contained in an early test, the *Hillbrand Sightsinging Test*, serve to illustrate them.² In this test, the pupil briefly studies each song presented and then sings without accompaniment. The following types of errors are noted:

1. Notes wrongly pitched
2. Transpositions
3. Times flatted
4. Times sharpened
5. Notes omitted
6. Errors in time
7. Extra notes

Our final group of specific dimensions in music concerns some of the aspects of instrumental performance. The exact appropriateness of each dimension, of course, depends upon the instrument being played.

Tone: Beauty and quality, intonation, fluency, and modulation.

Technical proficiency: Fingering, precision, intervals, breath support, tonguing, attack and release of tone, accuracy of notes and rhythm.

¹ New York, Psychological Corporation, 1939, and New York, Carl Fischer, Inc., 1930, respectively. For more complete data on standardized tests in music see Appendix B, page 482.

² Yonkers, N. Y., World Book Company, 1923.

Interpretation or musicianship: Phrasing, tempo, dynamics (shadings from soft to loud), rhythmic flow, expression, contrast, mood, naturalness, balance.

General effect: Sincerity, discipline, stage appearance, confidence, spirit, posture.

Industrial Arts. First, we shall indicate several of the dimensions of the working process in a shop class, and following this will be some of the important dimensions of shop products.

Shop performance:

- Ability to follow directions
- Care of materials
- Skill in handling tools and equipment
- Observance of safety precautions
- Adaptability when difficulties arise
- Ability to plan a procedure
- Ability to prepare a bill of materials
- Understanding of limitations and capabilities of tools and equipment

Shop products:

- Correspondence of finished product to original plans
- Neatness of over-all appearance
- Accuracy of angular measurements
- Suitability of finish
- Appropriateness of materials
- Fit of joints
- Accuracy of dimensions

Business Education. Performance in typing, shorthand, business machines, and filing (the activity aspects of business education) are measured largely in terms of the two primary dimensions of nearly any performance: speed and accuracy. Among the few special dimensions involved are:

Typing:

- Posture
- Hand position
- Stroke
- Paper insertion, margin setting, etc.
- Care of machine
- Appearance of copy
- Adherence to proper forms

Shorthand:

- Clarity of characters

Business machines:

- Setting up for given operation
- Checking or proofing
- Care of machine

Home Economics. The dimensions of performance in home economics have to do with all the various subdivisions of the subject: home management and family relationships, foods and nutrition, and clothing and textiles. Among them are the following:

Home management and family relations:

- Planning family responsibilities
- Budgeting
- Investment and financial planning
- Purchase of furniture and equipment
- Care of furniture and equipment
- Planning interiors
- Decorating
- Child care

Foods and nutrition:

- Meal planning
- Preparing and serving meals
- Purchasing and storing food
- Applying principles of nutrition

Clothing and textiles:

- Selection of clothing and household fabrics
- Making, repairing, and altering clothes (many detailed dimensions here)
- Laundering
- Use of sewing machine
- Choice and use of textures, color, and style in fabrics

Products in Home Economics are largely restricted to food prepared and garments made. Among the commonly measured dimensions for foods and clothing are:

Food preparations:

- | | |
|-----------------|------------------|
| Appearance | Texture |
| Consistency | Taste |
| Flavor | Tenderness |
| Color | Moisture content |
| Lightness | Odor |
| Greasiness | Arrangement |
| Size of serving | |

Garments:

- Attractiveness: Color and texture
- Choice of trim
- Pressing and cleanliness
- Style
- Individuality
- Suitability of style and fabric used

Workmanship: General construction
 Seams
 Fitting details (gathers, darts, pleats, tucks)
 Finishing details (collars, cuffs, waistbands, zippers, hems)
 Wearability

Physical Education. Performance in physical education involves general physical condition, general physical performance, fundamental athletic skills, and skill in specific games and sports.

Physical Condition. Various anthropomorphic measurements play an important part in physical education instruction. These include measures of such dimensions as height, weight, chest, stature, vital capacity, and respiration. These measures are often combined to establish various indexes such as build index, vital index, and ponderal index. These indexes, together with measures of motor performance, are used in the classification of students for equality of competition.

Physical Performance. Various aspects of physical performance are as important as physical condition in determining a pupil's achievement in physical education. Some of the more important dimensions of physical performance are:

Motor educability (the ability to learn new motor skills)	Form
Speed	Adaptability
Endurance	Co-ordination
Balance	Rhythm
Flexibility	Strength
Power	Agility

Athletic Skills. Certain fundamental athletic skills are believed to be basic to nearly all sports and athletic performances (exclusive of swimming). Among those which have been isolated are the following:

Running	Jumping
Throwing	Catching
Kicking	Pushing
Pulling	Dodging
Hand—eye co-ordination with an implement or bat	Strength of torso muscles
Moving quickly while carrying an object	Balance
	Ability to get over an obstacle
	Control of body in air or while hanging by arms

Sports Skills. Because of the great variety of sports, it is impossible in this text to list even a few of the special skills involved in each. Many instructional manuals have been published for the sports and games practiced in

the schools and these indicate their specific dimensions. Among them are found such unique things as *stroking out of a trap* (golf), *free throws* (basketball), and *sliding* (baseball).

This completes our enumeration of some of the specific dimensions involved in activity subjects. The identification of appropriate dimensions is the first step in any measurement process and the above lists, together with the discussion of general dimensions, should indicate the scope of the task.

Forms and Procedures for Measuring Performances and Products

Now that we have determined to some extent *what* is to be measured in the activity subjects, we must consider the form in which they may be measured and the procedures appropriate for measuring them. You may recall that three forms of symbolic expression may be used to characterize the status of a dimension. These are description-classification, ranking, and scaling. Both classificatory symbols and rank symbols may be assigned to any of the dimensions of performances and products. Description has particular use for such complex general dimensions as discrimination, timing, and coherency, and for many of the special ones: for example, composition and sympathy for subjects (art), interpretation (music), correspondence of product to plans (shop), budgeting (home economics), and co-ordination (physical education).

Because both performances and products have definite physical qualities, it is possible to use scale measurement somewhat more for them than for the dimensions of knowledge and understanding involved in other subjects. Speed and rate may, of course, be measured by time scales and such physical dimensions of products as are important may be measured in terms of feet, pounds, and color spectrums. In addition, error counts (accuracy) may be considered scale numbers if the errors may be assumed to have equal importance. Scale measurement is, of course, widely employed in physical education for the dimensions of physical condition and/or some of the dimensions of physical performance and athletic skills, e.g., lifting, running, jumping, throwing, and climbing. The applicability of scale measurement to sports is dependent upon the nature of the sport.

Observation and Product Analysis. Of the types of measuring procedure discussed in Section 1, observation seems to be the most appropriate for measuring performance, and product analysis is, of course, the natural procedure for use with products. The nature and proper use of observation and reliable ways of appraising products are discussed thoroughly in Chapters 4 and 5. The principles developed there are completely applicable to subjects now in question, and, moreover, many of the examples in these chapters have to do with art, industrial arts, physical education, etc. Consequently, it is thought unnecessary to repeat the discussion here. Instead, we should like to deal with a very specific technique of measurement that is especially relevant to the subjects we are now discussing. This is the "performance test."

Much of evaluation in the performance subjects is and should be based on observation of students at their regular work and on measurement of the things they produce in the ordinary course of instruction. If this is the only basis for evaluation, however, there is much room for error. Measures of different pupils are not always comparable, some aspects of achievement will unwittingly be stressed more than others, and some pupils will receive more attention than others. To supplement this evaluation through incidental observation, and to replace it at times, the use of a "performance test" is advisable.

THE PERFORMANCE TEST

A "performance test" is no more than a special instance of observation or product analysis planned so that given dimensions may be measured under given circumstances. Some task is specified that will require the pupil to engage in essential operations. The teacher watches him closely or scans his product closely and records some measure of his performance. This is usually in the form of a rating on some classification scheme, but it may be an indication of time, an error count, or simply a description of the performance. The advantage of a "performance test" over observation of ordinary performances and appraisal of ordinary products is that it yields measures more comparable from pupil to pupil and it permits the measurement of desired dimensions under controlled circumstances.

There are at least three basic types of "performance tests."³ The student may be required to:

1. Identify or recognize the proper procedure or the proper tools or parts in carrying out the performance.
2. Carry out the performance under simulated conditions or in miniature.
3. Carry out a single task that is typical of the over-all performance from which it was drawn.

Recognition Tests. In the recognition type of performance test, the student is confronted with a task given either orally or in writing and is asked to identify the proper procedure or the correct tool or piece of equipment to be used in performing it. In illustration, a student in a woodworking class might be asked to imagine himself faced with the task of cutting a groove of given dimensions in a piece of wood. He would then be asked to tell what tools he would use and how he would proceed. Another example would be in an electrical shop where the student is asked to identify the proper electrical wire splice from among several splices.

Since the student is not asked to carry out the actual performance, this type of test is at best an indirect procedure for measuring performance. The

³ These three types of performance tests follow the suggested classification presented by Ryans and Frederiksen (14:457-463).

recognition procedure is useful only in performances where several alternatives for tools and for procedures exist and where the actual manipulative movements are routine matters that everyone can do. The recognition type of performance test is particularly useful, then, for activities where choice or discrimination is the crucial dimension and where other aspects are relatively unimportant.

The Simulated Performance. The second type of performance test is the simulated performance. Here the student is asked essentially to "go through the motions" of a performance without actually carrying it out. Often this sort of test is used for measuring performances that involve expensive or inaccessible equipment. The armed forces, in particular, have made liberal use of the type. For instance, to observe the actual performance of a bombardier would require the use of an expensive bomber and the time of several crew members, to say nothing of gasoline, maintenance, etc. Consequently, the Air Force has developed mock equipment on which the bombardier simulates an actual bombing operation, and his performance is measured in this simulated situation.⁴

Basically, the simulated test requires the selection of the essential activities involved in a performance and then the provision of means for duplicating or simulating these activities where they may be easily observed. The effectiveness of the test is heavily dependent upon the degree to which the actual operation is simulated. Even at best, however, the simulated test contains some element of artificiality. It is likely that this test will always be used to some extent whenever expensive machinery, time, convenience, and safety are overriding considerations.

In schools, the use of the simulated test is not extensive, largely because performances involving expensive equipment are not a usual part of a school curriculum. The simulated situation is employed mostly in subjects where it is difficult to observe a performance closely as it actually occurs. Examples of this are afforded by "shadow boxing" and swimming out of water. In shadow boxing, the boxer contests an unseen opponent, thus allowing the instructor to observe footwork, use of arms, head, and body, and general form without the distraction of blows and the opponent. Likewise, if a swimmer lies on a low stool on his abdomen and then simulates a swimming stroke, the instructor can observe his co-ordination of arm, leg, and head movement much more exactly than he can when the student is in the water.

The Work Sample Test. The third type of performance test is called a "work sample" test. Here the student is asked to carry out a task typical of the over-all performance from which it was drawn. Of wide application and an extremely flexible procedure, this type of test has much to commend it

⁴ A model representing the earth's terrain in the vicinity of the bombing area is constructed. A radar screen using sonic waves instead of electrical waves depicts the earth's terrain as it moves over the model. Instruments simulate the movements of the plane and the bombardier uses this information to determine the bomb release point.

for use. The sample task may be carried out under actual conditions, hence the performer is more inclined to feel that his skill is being tested under realistic circumstances. If the sample has been carefully selected, the test may be a valid indicator of a student's ability to perform the activity as a whole.

The "work sample" is the type of performance test most commonly used in the schools. Therefore, the remainder of our discussion will deal with the use of work sample tests in the various performance subjects. Although such tests often are used to determine aptitude for certain activities, we shall concern ourselves exclusively with achievement. First, we shall outline some general principles that govern the construction of this type of performance test. Following this, we shall offer some examples of the tests as applied to specific activity courses.

Construction of Performance Tests of the Work Sample Type

The usual steps in constructing a performance test are to:

1. List all the specific activities in the performance that the test is to measure.
2. Select the activities that are to be included in the test.
3. Develop a task or series of tasks that incorporates these activities and manifests their dimensions.
4. Develop an observation form for measuring the activities in terms of their important dimensions.
5. Develop instructions, directions, and an over-all plan for administering the test.

JOB ANALYSIS

The first step is essentially a "job analysis." Here the teacher needs to jot down or at least to review mentally the significant activities that the performance involves. In selecting from this parent list the activities to be tested, the second step, the following criteria are suggested. The activities should:

1. Represent the whole performance as accurately as possible.
2. Be crucial in nature and have widespread effect on the quality of performance.
3. Reflect the emphasis given in instruction.
4. Embody the dimensions that meet the essential conditions of measurability.
5. Require minimal time and expense.

DESIGNING THE TASK

After the essential activities have been carefully selected, the third step is to design a task that incorporates these activities. This is just the reverse of what happens in an actual situation. Ordinarily, a task needs to be done and the activities occur in response to the requirements of the task. Here, how-

ever, we have a list of activities that we wish to observe and we develop a task that permits their observation. Consequently, the appearance of artificiality must be avoided and the task should be made as real as possible. This task, designed to incorporate the representative activities, is now the "work sample."

PREPARING THE OBSERVATION FORM

The fourth step, developing an observation form for recording measures of the work sample, is a crucial one. Before this form is developed there needs to be careful analysis and an isolation of the measurable dimensions of the activities involved in the work sample. The list of possible dimensions of performance suggested earlier in this chapter should be helpful in this analysis. For these dimensions or aspects to be measurable, they should be clearly defined and manifest easily observable variations. The observation form is no more than a list of the dimensions to be measured and a formal provision for recording variation relative to these dimensions.

Most performances involve stages or phases, which are their elemental dimensions. This is true in such performances as cooking, sewing, metal work, and mechanical drawing. In cooking, the stages might be the assembling of ingredients and equipment, the mixing of the ingredients, and the cooking. In mechanical drawing many phases may be identified, some of which are the planning stage, including freehand sketch and layout plan, the execution stage in which a pencil drawing is made, dimensioning and labeling, and finally inking. The observation forms for such performances should contain these stages or phases (the elemental dimensions). In addition, they must provide for an indication of the presence or absence or status of these elements. Usually the status of the elements is indicated with a reference to given properties or subdimensions of the elements, such as speed, accuracy, etc. In such case the observation form should provide a scheme for recording variation for each of these subdimensions.

Observation forms are classified according to the type of variation they record and the manner of recording it. The three basic types are check lists, descriptive or graphic rating scales, and anecdotal forms. Each type may be found in many different forms, depending upon the dimensions being measured and the nature of the activity being observed. The characteristics and uses of the three types are discussed fully in Chapter 4.

Check Lists. The check list form is particularly useful for measuring the accuracy aspect of performance. For instance, a teacher may prepare a list of procedural steps that should be followed. Each step is provided with a "yes-no" response on whether or not the step was satisfactorily performed. Accuracy of procedure is then directly measured by the total number of steps satisfactorily performed, or measured inversely by the number of errors committed. In other cases a check list is used of all the common errors of procedure or errors in following instructions, and the fewer the errors committed,

the more accurate is the performance. If a particular step or type of error is more important than others it should be weighted accordingly.

The following is an example of a check list type of observation form:

Check List for Pattern Cutting in Sewing

	Yes	No
1. Assembles necessary material and equipment	_____	_____
2. Lay out patterns on cloth for optimum use of cloth	_____	_____
3. Adequate pinning of patterns on cloth	_____	_____
4. Proper use of scissors	_____	_____
5. Cutouts closely approximate pattern shape	_____	_____
6. Proper allowances made in cutouts	_____	_____

Rating Scales. The second type of observation form is one that records several gradations of performance. In this form the "yes-no" or occurrence-nonoccurrence responses of the check list are replaced by descriptive levels or graphic ratings. In the check list form, just a check, a number, or the words, "yes-no" appear, and what constitutes satisfactory or unsatisfactory performance or the occurrence or nonoccurrence of an error usually is present only in the mind of the scorer. In the rating scale, on the other hand, a description of what constitutes each level of performance may appear on the form for everyone to see. This makes it possible for scorers to check one another more closely and for the persons being scored to find out more exactly the nature of their performance. In some "scales" these descriptions are omitted, gradations of performance being implied and symbolized by units of a line or by numbers in a series. This usually is a less satisfactory type than the descriptive scale.

A brief example of a descriptive scale for performance in throwing a pass in football is provided as follows:

Level I. Shows poor form. Ball falls short or overshoots. Not balanced and has little self-assurance.

Level II. Body is well balanced. Opposite foot points properly and hand is in proper place on the ball. Moves are somewhat jerky and self-conscious. Watches receiver too long. Ball travels in a wobbly manner but manages to get to receiver.

Level III. Body is well balanced. Opposite foot points to where ball is passed. Hand is in proper place on the ball. Moves with ease and assurance. Doesn't look directly at receiver until the proper moment. Ball travels spirally and is accurate.

In our later discussion of evaluation, it will be seen how easily the descriptive level type of observation form, such as the one above, can be adapted to making evaluations of performance.

Anecdotal Records. The third type of observation form is the anecdotal form. This is the most informal type of observation, since it consists essen-

tially of a blank sheet of paper on which the teacher records as objectively as possible what is observed in a performance. This type of form would be particularly useful when the performance situation is new to the teacher and somewhat fluid so that no clearly defined dimensions have as yet been established. Where description is to be the form of measurement, the anecdotal form obviously is the only appropriate recording device.

Many varieties of these three types of observation forms are either available or can be developed. Chapter 4 provides examples of the possible variations of rating scales. It cannot be emphasized too much that the development of an effective observation form is a very crucial step in the construction of a performance test.

PLAN FOR ADMINISTRATION

The fifth and final step in the construction of a performance test is to develop instructions, directions, and an over-all plan for administering the test. This part may be as formal or informal as the circumstance demands. In some situations with small groups, the test may be administered informally. A more formal situation, however, may be required for large groups. In some cases it may be desirable to provide for recording incidental observations or things that happened during a performance that might have bearing upon an evaluation of the performance.

Whether the performance test is administered under formal or informal conditions, certain requirements of good test administration should be met. The test situation needs to be standardized as carefully as possible with respect to time allotments or allowances, sequence, warm-up period, placement of tools, equipment, materials, the number of trials to be allowed, etc. There need to be clear-cut instructions so that the tasks are fully understood by the examinees. Furthermore, no variation in these instructions should be permitted, otherwise some examinees will be given an advantage over others. Also directions need to be provided for the person administering the test as to where he is to stand, how he is to behave, what he is to say, what specific observations he is to make, how he is to record his ratings, etc.

These instructions and directions comprise the over-all operating plan for administering the performance test. The plan should be written out when the test is complex or where large groups of students or several examiners are involved. If it is a simple test and few students and only one teacher are concerned, an outline or notes may suffice.

Examples of Observation Forms and Performance Tests Used in Specific Subjects

To conclude our discussion of measuring procedures in the performance-activity subjects, we want to present some examples of what teachers actually do when they measure pupil performances and products in art, music, shop, etc. Some of the examples are simple forms that may be used in incidental

observation as well as in performance testing. Others are performance tests per se. Since the devices illustrated were designed by teachers having only basic training in measurement, they are imperfect in some respects. However, it is thought that such "realistic examples" may be of more help to the beginner than would be idealized ones. Important errors will be indicated and improvements suggested.

ART

The form presented in Table 23 is typical of what might be designed by an art teacher to measure performance in a high school freehand drawing class where pencil and charcoal are used.

TABLE 23
Rating Chart for Art Performance
(Freehand Drawing Class)

	D	C	B	A
1. Drawing				
a. Accuracy of proportion or Suitability of distortion				
b. Relationship of proportions				
c. Stability of subjects				
d. Ease of interpretation				
2. Composition				
a. Balance				
b. Rhythm				
c. Spatial relations				
d. Textural interest				
3. Feel for Medium				
a. Line quality				
b. Tone quality				
4. Subject Matter				
a. Interest				
b. Arrangement				
<i>Key to Variations</i> D—Drawing shows no regard for aspect being judged. C—Aspect not well utilized. B—Aspect noteworthy, but room for improvement at this grade level. A—Aspect adds materially to the excellence of the picture.				

The rating chart in Table 23 is applicable to the *product* of art performance rather than the process. An excellent beginning has been made but the

levels of performance need to be more accurately described. Notice that general dimensions, drawing, composition, etc., are analyzed into their fundamental components.

DRIVING

Table 24 illustrates one way of rating "behind-the-wheel" performance in a driver training class. It is basically a check-list type of form, although it permits three-valued ratings.

TABLE 24
Check List for Driving Performance

<i>Aspect observed</i>	<i>Classifications of effectiveness</i>		
<i>Student posture</i>	YES	QUESTIONABLE	NO
Seat adjustment made	YES	QUESTIONABLE	NO
Mirror adjustment made	YES	QUESTIONABLE	NO
Foot position (dimmer switch and accelerator)	CORRECT	PAR. CORRECT	INCORRECT
Hand position (10 and 12 o'clock)	CORRECT	PAR. CORRECT	INCORRECT
Posture (erect and behind wheel)	CORRECT	PAR. CORRECT	INCORRECT
<i>Putting automobile in motion</i>			
Releases handbrake	YES	QUESTIONABLE	NO
Starts auto forward	SMOOTHLY	UNEVENLY	JERKILY
Shifts gears (low to second)	QUIETLY	SOME NOISE	GRINDING
Shifts gears (second to high)	QUIETLY	SOME NOISE	GRINDING
Steering in road	DIRECT	WEAVING	ASSISTANCE REQUIRED
<i>Bringing automobile to a stop</i>			
Puts hand out and down	PRECISE	UNDERSTAND- ABLE	UNIDENTI- FIABLE
Slows car down	SMOOTHLY	UNEVENLY	JERKILY
Brakes to a stop	SMOOTHLY	UNEVENLY	JERKILY
Sets hand brake	YES	QUESTIONABLE	NO
Parks car off pavement	YES	QUESTIONABLE	NO
<i>Showing consideration for others</i>			
When pulling out from curb	YES	QUESTIONABLE	NO
When stopping	YES	QUESTIONABLE	NO
Shows respect for rights of others	YES	QUESTIONABLE	NO
When in question as to others' rights, relinquishes his	YES	QUESTIONABLE	NO
Response to other drivers' signals and tolerant of their errors	YES	QUESTIONABLE	NO

MUSIC

The test described in Table 25 represents an attempt to measure performance in playing various musical rhythm patterns and is for use with beginning instrumental students. The test covers all the common rhythmic patterns through eighth notes and syncopation in 2/4, 3/4, and 4/4 time.

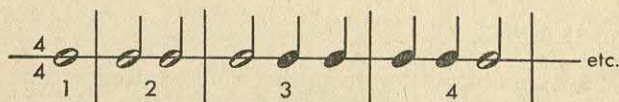
TABLE 25

Musical Performance Test

Rhythm Patterns: The rhythm patterns are grouped into five levels of difficulty, as follows:

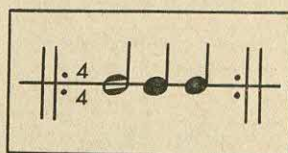
- I. All possible combinations of ♩ , ♩ , and ♩ .
- II. All possible combinations of ♩ , ♩ , and ♩ .
- III. All possible combinations of ♩ , ♩ , and ♩ .
- IV. All possible combinations of ♩ and ♩ .
- V. All of the above combinations plus syncopated patterns.

For example, at the first level we have the following possibilities in 4/4 time without using the syncopated pattern.

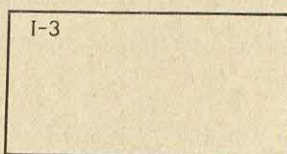


Each pattern is written on a large card, on the back of which is recorded the level of difficulty and the pattern number as shown below.

Example:



Front



Back

TABLE 25 (Continued)

Administration of test: The teacher places one of the cards on the music rack, establishes a tempo by tapping out one measure, and the student plays the pattern on his instrument at some convenient pitch, repeating the pattern several times. Each performance is rated according to the following variation scheme:

Rating	Description of Performance
1	Plays correctly on first attempt without hesitation.
2	Plays correctly on first attempt, but after first analyzing or thinking it through.
3	Plays correctly on second attempt.
4	Plays correctly after some stumbling and fumbling.
5	Is unable to play the pattern correctly.

The teacher records each performance by identifying the pattern number and the rating for its performance, as follows:

Pattern Number	I-4			
Rating	2			

This performance test attempts to measure two dimensions simultaneously, first, the level of complexity of the pattern and, second, how quickly the pattern was played accurately. Since the dimensions have been clearly identified and their variations carefully defined, the ratings on this performance test should be reliable. Because it is to be individually administered, the test would probably consume considerable time unless some sampling scheme were used.

INDUSTRIAL ARTS

In Table 26 is an observation form that may be used to rate a shop performance, applying an oil varnish. This form illustrates the measurement of a performance in which detailed procedural steps are given the most attention. Since some of the steps may be more crucial to the final result than others, these should be given additional weights if an over-all score is to be given to the student's work. An unsatisfactory aspect of the form is that it provides for immediate evaluation of procedures rather than their *measurement* and yet does not define what is a satisfactory as against an unsatisfactory condition.

TYPING

The plan in Table 27 is used to rate a student's ability to set up an outline in proper form. Students are given copy and told to type it according to specific written directions.

TABLE 26
Rating Form for Applying Varnish

<i>Action observed</i>	<i>Unsatisfactory</i>	<i>Partly satisfactory</i>	<i>Satisfactory</i>
I. Preparing surface 1. Checks dryness 2. Removes dust, using suitable cloth 3. Removes grease or wax II. Getting the varnish ready 1. Pours only enough varnish for job 2. Does not pour varnish back into can 3. Checks varnish flow and takes corrective steps if necessary III. Applying varnish to wood 1. Checks room temperature and ventilation 2. Sprinkles floor to lay dust 3. Checks clothing for dust 4. Selects brush of suitable size 5. Checks brush for cleanness and loose bristles 6. Dips brush into varnish about $\frac{1}{3}$ the length of the bristles 7. Taps brush lightly inside of container 8. Flows varnish on surface well 9. Starts at center of surface and brushes out toward the edges 10. Wipes excess varnish from the brush over the edge of the container 11. Evens up the surface with light feathering strokes 12. Works with grain			

TABLE 27
Plan for Judging a Typing Performance

-
1. *Neatness*—20 points
 Points will be subtracted if the following requirements are not met:
 - a. Even touch to stroking of keys to avoid light and dark letters
 - b. No strike-overs
 - c. Clean appearance of paper with no smudges, finger marks, creases, etc.
 - d. Good general appearance
 2. *Accuracy*—50 points
 - a. Typographical errors will be penalized on the basis of two points off for every error
 - b. Total points subtracted for typing errors shall not exceed 50

TABLE 27 (Continued)

-
-
3. *Directions Followed*—20 points
 - a. Use of plain $8\frac{1}{2}$ " x 11" paper
 - b. Right number of spaces for top margin
 - c. Correct placement and capitalization of heading
 - d. Proper number of spaces between heading and first line as well as correct spacing between lines for the remainder of the outline
 - e. Indentations properly placed in five spaces
 - f. Correct number of spaces after all periods
 - g. No strike-overs
 - h. No erasures
 4. *Completion of Exercise*—10 points
 - a. Ten points allowed if entirely completed
 - b. Adjust this part of the score commensurately with amount finished
-

The reliability of this type of point rating is low because of the subjective basis for assignment of points to dimensions 1 and 3. However, the form does show a good operational analysis of the factors in a typing performance.

HOME ECONOMICS

The rating outline in Table 28 is used in a home economics class and is similar to the one for typing. The dimensions of a product are broken down and are to be assigned points on a more or less subjective basis. As with the

TABLE 28

Outline for Rating a Gathered Skirt

(Each of the eight groups is worth 5 points.)

1. Suitability of material to the type of skirt.
2. Direction of cut suitable to the pattern of the material.
3. Seam construction
 - a. Seams straight
 - b. Threads tied
 - c. Proper width ($\frac{1}{2}$ "– $\frac{5}{8}$ ")
4. Side opening.

Zipper:

 - a. Lies flat
 - b. Stitching straight
 - c. Opens and closes easily

Continuous lap:

 - a. Proper width in relation to weight of material
 - b. Lies smoothly
 - c. Not wrinkled
5. Gathering threads sewed in evenly and pulled up evenly.

TABLE 28 (Continued)

6. Belt
a. Not stretched to one side or other
b. Stitched evenly
c. Buttonholes centered
d. Buttons sewed securely, fastened neatly
7. Hem
a. Lies smoothly
b. Even width
c. Stitched neatly and securely
8. Fit of skirt
a. Waist band proper size
b. Skirt length correct
c. Skirt hangs evenly

typing scale, the basis for assignment of points is too indeterminate. Just how would a hem look to be rated 4 points rather than 3? The form would be improved by inclusion of described gradations that are worth so many points.

PHYSICAL EDUCATION

A plan for evaluating a dive is shown in Table 29. Any dive has been broken down into five components, with a description of what constitutes good procedure for each. The dive is rated at one of five levels according to the diver's conformance to the prescribed procedures.

TABLE 29
Plan for Evaluating a Diving Performance

Aspects of Diving Performance
Readiness:
Balanced, erect, well poised
Approach:
Steady, easy, smooth, erect, legal number of steps
Takeoff:
Proper position on board, jump is timed with board, graceful and effective use of arm and leg movement, body erect
Flight:
Sufficient height for necessary movements, body movements conform to specifications of dive (jackknife, twist, layout, fold, turns), movements are smooth, easy, and graceful

TABLE 29 (Continued)

Water entry:

Body perfectly straight with feet together, in approximate line with diving board, minimum splash and sound

Rating Scale for Over-all Diving Performance

Level	Description
1	Dive is only barely recognizable. Basic errors committed throughout. No apparent control.
2	Some control is apparent but dive is still inadequate. Many errors made.
3	Minimum requirements of dive are met but movements are still somewhat jerky. Lacks control in some respects.
4	Major requirements are met with general impression of good control and form, some minor errors and deviations still exist.
5	All aspects of dive skillfully and gracefully executed with no apparent variations.

The scale could be improved by relating each level explicitly to each of the components of a dive. As it is, any rating is a quick and subjective "average" of performance for each of the components.

SPEECH

While the measurement of speech is treated in Chapter 10, speaking is an activity subject and it affords an example of a novel and carefully worked out performance test. The test outlined in Table 30 is designed for use in the primary grades and is to be administered individually.

TABLE 30

A Picture Articulation Test for Nonreaders

Errors in articulation consist of the following three types:

1. *Omission* of sound:
 - a. Example: says "ca" for "cat"
 - b. Record error as follows: o/t (omits *t*)
2. *Distortion* of sound:
 - a. Example: an *s* that is whistled (sound is sloppy or inaccurate)
 - b. Record error as follows: /s (the *s* is distorted)
3. *Substitution* of sound:
 - a. Example: says "wabbit" for "rabbit"
 - b. Record error as follows: w/r (substitutes *w* for *r*)

The place of the articulation error in the word mispronounced is usually indicated as follows:

- I—Initial position
 M—Medial position
 F—Final position

TABLE 30 (Continued)

Examples:

th/s(F)—substituted *th* for *s* in the final position of the word.

/s(I)—distorted the *s* sound at the beginning of the word.

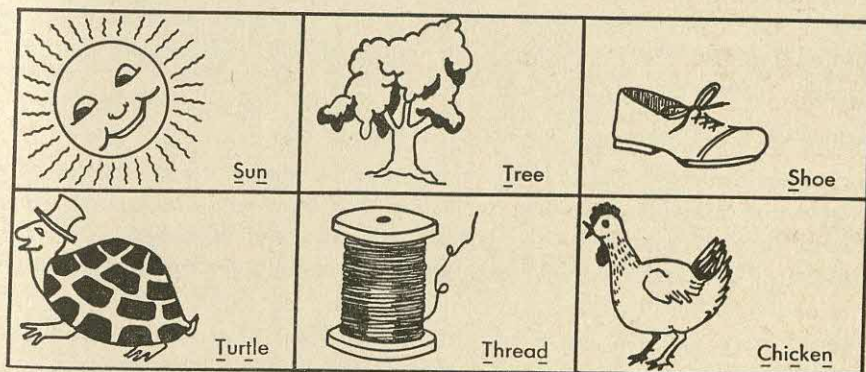
o/s(M)—omitted the *s* sound in the middle of the word.

Procedure for administering the test:

1. Engage the child in spontaneous conversation and record articulation errors that are noticeable in this type of speaking.
2. Show child the picture cards in sequence. Do not say the word for the child.
3. Rate the child on the rating scale of severity.

Picture Cards:

The following is a sample of the picture cards used to elicit the necessary sounds. Alongside each picture is the word that the picture should elicit, and underlined are the sounds being observed. The words, of course, would not appear on the actual picture.



Rating scale of severity of misarticulation:

Directions: Find the category that most nearly applies to the child's speech and record the category number. If two categories seem to apply, record both category numbers and underline the one that seems to be slightly more pertinent to the case.

Example: 2-3 would mean that the child is on the borderline between No. 2 and No. 3 categories, with a slight edge given to the latter.

Category

Description of Misarticulation

- 1 Mild, but noticeable defect. May be considered normal unless accompanied by psychological or etiological factors. Errors in 1 or 2 sounds.
- 2 Moderate disability. Noticeable deviation, but does not interfere with normal communication. Prognosis: good. Probably amenable to brief speech therapy.
- 3 Obvious articulation defect. Errors in three or more sounds. Gives evidence that speech disability has become habitual and will require special therapy for re-education. Immediate therapy needed.
- 4 Severe articulation problem. Immediate therapy needed. Speech is so noticeably deviating from normal that the child suffers psychologically in his classroom or at home. Prognosis may be poor or good for improvement.
- 5 Extremely severe defect. Immediate need for therapy. Speech indicates need for intensive extended therapy. Prognosis may be poor.

All the essential phases in performance testing are embodied in this test, namely: an identification of the dimensions of the performance, development of a task that provides for observation of the dimensions, development of a rating scale and a recording system, and finally an over-all plan for administration of the test. The rating scale in particular can be improved by replacing such qualitative words as "mild," "moderate," "obvious," "severe," and "extremely severe" by a more exact description of the actual nature of the articulation error.

This completes our illustrations of specific observation forms and performance tests. It is hoped that the examples are sufficient to suggest what is involved in their development. By now it should be clear that a great deal of thought and experience must go into the construction of an adequate observation form and/or performance test. Unfortunately, the work that is involved has discouraged many teachers. However, when one considers the alternative of depending upon subjective and unsystematic appraisals and the negative effect the latter may have upon pupils and instructional efficiency, it seems that any time and effort spent on developing valid observation forms and performance tests is entirely justified.

Evaluative Standards and Problems of Reporting

In general, evaluation of either performance or products should be governed by the principles described in Chapter 9. There should be a comparison between the performance or product and a known and appropriate standard for either. The standard should not only take into account the intrinsic character of the action or artifact but also the age of the pupils and the purpose of the course; that is, general education, vocational preparation, recreation, or whatever. The standard should express clearly defined levels or gradations of quality for each of the dimensions to be evaluated. Evaluative symbols assigned to the performance or product should be clearly related to the standard. Periodic reports on progress that cover many performances and products should be based on comprehensive measurement, be clear both to pupils and to parents, be somewhat diagnostic, and reflect consistency as to standard both among teachers and for the same teacher from year to year.

A basic consideration in evaluating achievement in such subjects as music, home economics, etc., is that the dimensions of performances often have immediate implications for the value of the performance. As there is increase in such general dimensions as accuracy, speed, discrimination, economy of effort, and coherency, there usually is an increase in the worth of the performance. And as there is decrease with regard to these dimensions the value of the process typically decreases. Moreover, where there are prescribed elements or steps (the elemental dimensions) in a product or activity, their inclusion generally adds to the value of the performance and their omission subtracts from its worth. For example, in applying paint it is necessary first to

prepare the surface by cleaning and/or sanding. A check that this has been done means that the painting task has been done so much the better.

For this reason an evaluative standard frequently is incorporated in the rating scales and other observation forms used in the performance subjects and hence evaluation and measurement are accomplished in a single act. This was characteristic of several of the examples of observation forms we just presented. Notice in particular the Check List for Driving Performance, the Rating Form for Diving, and A Picture Articulation Test for Nonreaders. Such union of measurement and evaluation is efficient and desirable as long as it is a planned procedure and as long as the appraisal is based on observation of the specific performance or product and not on the teacher's general feelings about the pupil. In evaluating, somewhat more than in measuring, it is relatively easy for a teacher to let his over-all opinion of the pupil affect his opinion of a given performance.

Earlier we asserted that perhaps the best type of observation and product analysis form was a descriptive or graphic rating scale. The scale contains brief descriptions of different grades or levels of performance. The pupil's performance is measured by noting which of the levels his behavior approximates and by assigning to him the symbol that represents that level.

Such a rating scale lends itself readily to combined measurement and evaluation. All that is required is that the value of each gradation of performance recorded on the scale be established and a symbol used that signifies this value. When the levels in the scale have been assigned their appropriate values, the scale then becomes an evaluative standard as well as a measuring device.

Although there may be a close affinity between the status of a performance with respect to certain dimensions and the value of that performance, there is no automatic correspondence. Appropriate assignment of values to different levels of performance must take into account the age of the pupils, the amount of instruction given them, the relationship of the given course to earlier and later courses, etc.

There are many publications to assist the teacher in defining valid standards for performance-activity subjects. Methods textbooks in the several subjects contain certain standards and provide the basis for devising others. Courses of study often contain them and many have been prepared and published by professional associations and committees representing given subject areas. The teacher should, of course, not depend solely on books for his standards. He should accumulate appraisals of typical performances and products over a period of time and continually study the capability and motivations of his pupils.

In addition to these general considerations, evaluation in several of the performance subjects involves special problems. Some of these are discussed in the following paragraphs.

Art. Three factors seem to constitute the principal difficulties in art evaluation and these must be circumvented or reconciled in some way if art evaluation is to be effective.

1. Excellence in drawing, painting, sculpture, etc., seems to necessitate special talent. Students in art courses differ widely in its possession and consequently any comparative grading or any use of absolute standards is certain to make for invidious evaluations.

2. All their lives children have viewed the stereotyped art of magazines, greeting cards, comic strips, textbook illustrations, advertising, and cartoons. Necessarily, their efforts at drawing and painting may imitate these. Many art teachers, on the other hand, have been taught to avoid stereotypes and to seek for original expression. Consequently, some pupils who are good craftsmen but are too much influenced by stereotypes may be judged adversely, while some "original" spirits may be commended despite many errors in technique.

3. There are several "schools" of art, each with its own philosophy, style, media, and subject matter. Some teachers are prone to teach in terms of a "school" and to evaluate by its standards. Pupils who are taught in successive classes by teachers representing divergent art philosophies may receive what they consider to be unfair evaluations in the second class.

Music. In music, the musical composition as written constitutes a sort of ultimate standard for any singing or instrumental performance—together, of course, with the accepted conventions of tone, rhythm, harmony, etc. As a student approximates the "music" as symbolized on the score his performance is good, and as he departs from it the performance is bad. For such an external standard to be used with fairness in public school music classes, it must be tempered by the teacher's knowledge of his students' prior training, their age, and their talent, etc. As in art, special talent is a variable that makes evaluation particularly difficult.

The tape recorder makes it possible for any music teacher to develop performance scales that may be used as evaluative standards. Many pupils judged as bad to fair to excellent performers can have their singing or playing recorded. From these many recordings a series can be selected ranging from best to worst and each may be assigned a given value. The performance of any pupil can be judged then by having him sing or play the same composition as is used in the scale.

Business Subjects. Evaluation of student performance in typing, shorthand, machine operation, clerical practice, and bookkeeping is relatively free from the problem of subjectivity that besets art and, to a lesser extent, music. It does, though, involve two other problems. For one, business courses are almost necessarily vocational courses or at least have immediate vocational significance and thus the standards of business practice must be considered. For the other, in many instances low-ability high school students are assigned to business classes simply because they cannot learn the more "academic"

subjects. Needless to say, this sort of educational guidance is of dubious value.

Employers want stenographers who can type and take dictation at given speeds and who can conform to certain conventions of letter composition, filing, telephone usage, etc. If the teacher is to be realistic, he must relate the standards in his class to those of the employers.

He can do this by establishing a single level or type of performance that is acceptable in the business world and then by evaluating performance as satisfactory that achieves the level and as unsatisfactory that does not achieve it. Or he can devise a performance scale containing several levels of competence, one of which is acceptable for business practice and is labeled as such. Obviously, from our point of view, the use of the performance scale is preferred. In beginning classes, particularly for eighth and ninth graders, the minimum business standard might be at the top of the scale or even beyond it. Business textbooks and teachers' guides usually are prepared in relation to business standards and most of their procedures conform to business practice. Hence, if course standards are derived from textbooks and guides they are likely to incorporate vocational standards.

The presence of low-ability students who have no interest in the subjects per se poses a dilemma for the business teacher. From one point of view, they are not likely to attain acceptable vocational performance levels and hence should not be allowed to pass. From another, they have to take the course, they certainly will learn something, and hence should be allowed to pass. Some resolution of the dilemma may be attained by use of a dual marking system, one standard relating to absolute achievement and the other to learning or progress.

Industrial Arts. Problems of evaluation in the various craft and shop courses—drafting, wood shop, machine shop, auto mechanics, etc.—are similar to those in business education. Instruction in this area necessarily has vocational significance, even in general shop and in nonvocational programs, simply because the things boys learn to do are what men do for wages. Even more than in business courses is there a tendency for trouble-makers and low-ability students to be assigned to industrial arts classes. As with business education, evaluative standards for “shop” courses must be related to vocational standards but they should also take into account the spread of achievement likely to occur in any class. The use of two standards, one for performance per se and the other for learning, and the assignment of marks on the basis of either, seems to be the best approach to fair evaluation of the students who are in a class for reasons other than interest and/or aptitude.

Physical Education. The male physical education instructor in a secondary school is usually the coach of one or more interscholastic sports and much of the instruction in physical education relates to these sports. This makes for one problem in evaluation. A second source of difficulty in evaluation is the large size of the usual physical education class and the routines of lockers, gym dress, and showers involved.

Because of the dual role of the coach and the "varsity" significance of many sports to be played and learned in boys' physical education classes, evaluative practices are susceptible to two inequities. The teacher may be tempted to use the performance of his letter winners as the basis for grading and thus give low marks to a large number of students. Or he may become preoccupied with judging the skill of the better athletes, those who are team prospects, and simply not evaluate the others, "giving" them a *C* for attendance and obedience. These errors in evaluation may be minimized, if not avoided, by the conscious and planned use of written or graphic performance standards appropriate to the spread of ability in regular classes, not just team classes.

With large classes, expensive equipment, special dress and sanitary routines, a certain amount of regimented behavior is mandatory in physical education classes. If the teacher permits too many individuals to deviate from the prescribed order, chaos may result and teaching suffers. So the need to have pupils conform becomes acute and one of the few "motivators" left to the modern teacher is the report card. Consequently, many physical education teachers reward consistent attendance, suiting-up, showering, replacement of equipment, etc., with increased marks and punish inconsistency in these affairs with lowered marks.

While without question this practice confuses the significance of physical education grades, its value for class control makes it very attractive to teachers. It is hoped, of course, that the necessary conformity can be gained without the involvement of marks in achievement. If it cannot, and the marks are to be affected by promptness in dressing, etc., pupils should know just how much weight is to be given the conformance factors and this weight should not be excessive. An evaluation of the pupil's physical and athletic achievement *per se* should be clearly indicated to him and his parents by a separate communication if necessary. Rather than lowering or raising an achievement grade, it seems better to have a separate mark for conformity to routines in physical education and, if necessary, to have this separate mark bear on promotion, honors, and even graduation.

Summary

This chapter was devoted to the measurement of the performance aspects of the school curriculum that constitute the major portion of such subjects as art, music, shop, driver training, home economics, and physical education. In the measurement of performance, it is helpful to identify two phases: process and product. Most performance processes involve in common the dimensions of speed, accuracy, discrimination, economy of effort, timing, intensity, and coherency. The measurable dimensions of performance products are their physical properties and such elements and qualitative attributes as derive from their specifications and purposes.

The status of performances and products may be indicated by classifica-

tion and descriptive symbols, by rank numbers, and, for some dimensions, by scale numbers. Observation and product analysis obviously are the basic measuring procedures to be used in the performance subjects. In addition to casual observation of processes and appraisal of products, it is well to use the device of the performance test. Such a test, in general, calls upon the student to perform or to simulate performing an activity under prescribed conditions. The three basic types of performance test are: recognition, simulated task, and work sample. In the preparation of a performance test, five basic steps are involved: itemize the specific activities to be covered, select those activities to be included in the test, develop a task or series of tasks that permit observation of these activities, develop an observation form for appraising performance and/or product, and, finally, develop an over-all plan for administering the test. The most crucial part is, of course, the observation form. Some illustrations of performance tests and observations forms were presented and their shortcomings discussed.

Evaluation of performances may be accomplished automatically if an evaluative standard is incorporated into the rating form. This requires that differential values be properly assigned to the gradations of performance described on the form. Among the problems encountered in evaluating performance are the importance of special talent in art and music, the influence of interscholastic competition in physical education, and the sometime use of business and shop classes as havens for slow or maladjusted pupils.

EXERCISES

1. Contrast the problems of measuring and evaluating in "activity courses" with those of measuring and evaluating in "academic courses." In what ways are the problems similar?
2. Select a standardized performance test and critically analyze it in terms of the dimensions being measured.
3. What would you reply to a teacher of music or art who maintains that measurement is impossible in art or in music because performances in these areas are so subjective?
4. Select a specific performance in your teaching area. Develop a set of evaluative standards and construct a device for measuring the dimensions of this performance.
5. How would you proceed systematically to improve your performance tests?
6. Suppose you were called upon to rate a performance with which you have had little experience. How would you prepare yourself for this task?

BIBLIOGRAPHY

1. Adkins, Dorothy, *Construction and Analysis of Achievement Tests*. U.S. Civil Service Commission, 1947, chap. 5, pp. 211-265.

2. American Association for Health, Physical Education, and Recreation. *Measurement and Evaluation Materials in Health, Physical Education and Recreation*. Washington, D.C.: 1950.
3. Army, Clara M., *Evaluation in Home Economics*. New York: Appleton-Century-Crofts, Inc., 1953, 378 pp.
4. Bovard, J. F., F. W. Cozens, and E. P. Hagman, *Tests and Measurements in Physical Education*, 3d ed. Philadelphia: W. B. Saunders Co., 1949.
5. Fleck, H. C., *How to Evaluate Students*. Bloomington, Ill.: McKnight and McKnight Co., 1953.
6. Glassow, R. B., and M. R. Broer, *Measuring Achievement in Education*. Philadelphia: W. B. Saunders Co., 1939.
7. Jordan, A. M., *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953.
8. Larson, L. A., and R. D. Yocom, *Measurement and Evaluation in Physical, Health, and Recreation Education*. St. Louis: C. V. Mosby Co., 1951.
9. Malraux, André, *The Psychology of Art* (translated by Stuart Gilbert). New York: Pantheon Books, 1949.
10. Michaels, W. J., and M. R. Karnes, *Measuring Educational Achievement*. New York: McGraw-Hill Book Co., Inc., 1950, chap. 12.
11. National Society for the Study of Education, *The Forty-Fifth Yearbook*, Part I, "The Measurement of Understanding." Chicago: University of Chicago Press, 1946.
12. Newkirk, L. V., and H. A. Greene, *Tests and Measurements in Industrial Education*. New York: John Wiley & Sons, Inc., 1935.
13. Patterson, C. H., "Predicting Success in Trade and Vocational School Courses: Review of the Literature," *Educational and Psychological Measurement*, 16:352-400, 1956.
14. Ryans, D. G., and N. Frederiksen, "Performance Tests of Educational Achievement," *Educational Measurement* (E. F. Lindquist, ed.). American Council on Education, Washington, D.C., 1950, chap. 12, pp. 455-494.
15. Seashore, E. C., *Psychology of Music*. New York: McGraw-Hill Book Co., Inc., 1953.
16. Schoen, Max, *The Psychology of Music*. New York: The Ronald Press Co., 1940.
17. Scott, M. G., and E. French, *Evaluation in Physical Education*. St. Louis: C. V. Mosby Co., 1950.

CHAPTER 14

INTELLIGENCE

In the preceding four chapters we have studied the evaluation of pupil achievement in various school subjects and the need for such evaluation was obvious: the pupil must see how he is progressing and the teachers must see what they are accomplishing. Now, however, we are about to be concerned with the measurement of something that isn't a school subject, *intelligence*. And it seems appropriate to ask why we should study its measurement.

Intelligence Defined

According to American usage as recorded in dictionaries, intelligence is “. . . the capacity for knowledge and understanding, especially, as applied to the handling of novel situations; the power of meeting a novel situation successfully by adjusting one's behavior to the total situation; . . .¹ Apparently, the idea of intelligence is very old and widespread. Plato in his *Republic* wished people to be classified for vocations on the basis of mental differences. His philosophers (lovers of wisdom) were to be the rulers. So far as is known, all languages have a word for it and all ages and kinds of people are purported to have some of it. We say, “What an *intelligent* child” and “the Bushman has less intelligence than we.” Furthermore, by intelligence apparently is meant something that the brain does or is. Words like cerebration and gray matter, phrases like “What a brain!” tend to recur as intelligence is discussed. Finally, we seem to mean something that changes or grows when we speak of a child “becoming more intelligent” and when we use mental age as an index of intelligence.

What sort of thing is intelligence then?—to be a power or act, to be old and universal, to have to do with the brain, to grow, and, particularly, to have a thousand or more definitions. In this last characteristic may be the key to its nature. *Intelligence* apparently is an *explanation*: an explanation for the differences we observe among people in their capacity for learning, in their remembering, in their problem solving, etc. Since an explanation is never seen or touched, it is easy to say different things each time we attempt the explanation. So it is hardly surprising that intelligence is given so many varied meanings.

¹ *Webster's New International Dictionary*, Second Edition, Unabridged, Springfield, Mass., G. C. Merriam Company, 1952.

Intelligence is that type of an explanation that deserves the term *construct*. You may recall the discussion of constructs in Chapter 2 (pages 26–28). They are the verbal or mathematical “maps” that enable men to deal more rationally with, and often to control, complex observable phenomena. The complex observable phenomena in this case are a very large number of behaviors that seem to have something in common. Examples of these behaviors are reading, figuring, searching, writing verse, learning a new language, and solving mechanical problems. The construct of *intelligence* enables us to understand individual differences in such behaviors, permits us to appraise their relation to one another, and allows us to predict the achievement of given children in reading, history, chemistry, etc.

Thus, it seems that the measurement of intelligence needs to be discussed, even though intelligence is not a subject of instruction, because it affects pupil achievement in subjects that are taught. In our presentation, the first consideration is the dimensions of intelligence for which measurement is attempted; next, the customary forms and procedures for its measurement. In connection with the latter, attention is given to the validity and reliability of intelligence tests, to IQ variability, and to the practical uses of intelligence testing in school practice.

Dimensions of Intelligence²

The dimensions of intelligence are, of course, the properties or variables imputed to the construct toward which appraisal is directed. There seem to be three general sources of them: theories of intelligence, the nature of the tests, and vernacular usage.

THEORETICAL VIEWPOINTS

One theorist, Thurstone, has analyzed the results of various kinds of intelligence tests mathematically to determine what the dimensions of intelligence are. His method is called Factor Analysis and it has yielded certain “primary mental abilities”: spatial, perceptual, numerical, verbal relations, memory, words, induction, reasoning, and deduction (43). Another theorist, Stoddard, has used observation and logic as his principal tools of investigation and, from a different point of view, offers quite different dimensions. As he puts it, “Intelligence is the ability to undertake activities that are characterized by (certain properties) . . .” The properties, or, as we would say, dimensions that he asserts for these intelligent activities are: “1. difficulty, 2. complexity, 3. abstractness, 4. economy, 5. adaptiveness to a goal, 6. social value, and 7. the emergence of originals . . .” (40).

Somewhat contrary to the viewpoints of these two who differentiate numerous dimensions are the opinions of two other imposing figures in the

² This treatment of dimensions is cursory and for purposes of general understanding only.

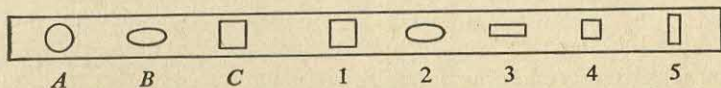
field of intelligence testing. Spearman, an English psychologist, asserts that two factors contribute to every intelligent behavior (38). One, a general factor called simply *g*, functions in all situations. In any given situation an additional specific factor, *s*, is also operative. Binet, the "father" of intelligence testing, seemed to think that intelligence was more or less a whole or integral aspect of the individual, although he was not given to detailed definitions (4).

DIMENSIONS INHERENT IN TESTS

The questions and problems of intelligence tests, our second source, permit us to infer several dimensions that resemble these of theorists and others that are of a different order. One very usual type of item in primary level tests requires that the child differentiate among a series of animals, household articles, geometric forms, etc. Such a task requires that the child *perceive discriminately* and, hence, a dimension of *discrimination* is implied.

In the 1937 Stanford Revision of the Binet (Appendix B, page 478), children are asked to look at two images for ten seconds and then to reproduce what they saw. Again in the Binet and in the Wechsler-Bellevue (Appendix B, page 479), testees hear a series of digits and are required to repeat them just as they were said. Both of these tasks require that the child remember or recall what he just perceived. Thus, *recall* may be inferred as a dimension that the test measures.

A somewhat different dimension apparently is measured by the myriad of analogies items which occur in intelligence tests. These are verbal, graphic, and numerical. An example of the graphic variety is found in the American Council on Education Psychological Examination, the so-called ACE.



(Educational Testings Service, Princeton. Reprinted by permission of Educational Testing Service.)

Here the student is asked to complete the proposition, "C is to _____ as A is to B," or, as the test directions suggest, "What will figure C be if you change it by the same *rule* as A was changed to make B?" It is obvious (and from your own test-taking experience you may recall) how similar verbal and numerical analogies go. The dimension appraised in these items apparently has to do with perceiving something in one situation and applying it to another. Let us call the dimension *generalization*. Items other than analogy types seem to measure this dimension. For example, one of the questions in the Otis Self-Administering Tests of Mental Ability (34) is

"A lion most resembles a _____"

1. dog, 2. goat, 3. cat, 4. cow, 5. horse."³

³ Copyright by World Book Company and reproduced by special permission.

In many tests children are asked to recognize objects that they have seen before and to name objects and pictures. Such items are called recognition and/or naming questions, and it may be that a correlative dimension of intelligence is being measured. This dimension (let us call it *recognition*), is akin to *recall* and *discrimination*.

Problem solution is an activity found in most intelligence tests and consequently deserves designation as a dimension of intelligence to be inferred from the tests. Most of the "problems" are mathematical but some are verbal (34): "If the following words were arranged to make a good sentence, with what letter would the second word begin? same, means, small, little, the, as."³ Still others are spatial. For example, in the Binet, Form L, the first test at age thirteen consists of verbal directions to draw the path you would take in searching if your wallet were lost somewhere in a round field (Appendix B, page 478).

In final illustration of dimensions that may be induced from the items of intelligence tests, let us take notice of the hosts of items that ask what does this mean, pick out the word most like this one, tell me about an "orange," etc. All such items require that testees give the same standard verbal responses to a given verbal stimulus. We call these standard verbal responses definitions, and some of nearly all intelligence tests and nearly all of some tests ask the testee to define words. Consequently, it seems appropriate to specify it as a separable dimension. *Definition*, then, is another of the dimensions of intelligence that we may infer from tests of intelligence.

VERNACULAR DIMENSIONS OF INTELLIGENCE

The third source of dimensions of intelligence provides items that resemble those of Stoddard and other "logical" theorists somewhat more than they resemble the items evolved from statistical manipulations. When parents try to characterize more intelligent children (usually their own), they use such terms as "alert, he learns quickly, thinks fast, and sensible." Educators have been talking for many years about gifted children and how to educate them. According to them, gifted (very intelligent) children are more critical, creative, given to abstractions, perceptive, imaginative, etc. Dimensions of this type are a good deal less well defined and far less mutually consistent than are those to be inferred from tests or stated by theorists. They have the advantage, however, of describing what the majority of Americans mean by intelligence, and any group of intellectual dimensions given general distribution must connote essentially what they connote if intelligence as measured is to mean intelligence as understood.

DIMENSIONS FOR WHICH THERE IS GENERAL CONSENSUS

From this brief review of the dimensions of intelligence offered by each of three sources, theories, tests, and ordinary usage of the term, it is apparent that there is no single, agreed-upon enumeration of dimensions for intelligence.

While the quip, "intelligence is what intelligence tests measure," represents an extreme viewpoint, it is well to consider that the "intelligence" measured by any given test is defined by the dimensions that the test attempts to measure. Fortunately, most test designers essentially agree in their choice of types of items, and, hence, their tests appraise much the same things. Some tests, of course, seem to include more dimensions than others but few are concerned with any "unique" property.⁴

The following list is thought to subsume most of the dimensions for which we could expect reasonable consensus among theorists, test designers, and professional users of intelligence tests and their results:

Recall (immediate and delayed)

Discrimination (likenesses and differences, details, patterns, relationships, etc.)

Verbal symbolization (naming, definition, predication, etc.)

Number symbolization (quantification, counting, arithmetic manipulations, etc.)

Abstraction (both verbal and numerical, inductive and deductive reasoning, classification, rule stating, etc.)

Invention (problem solving in verbal, numerical and mechanical situations, storytelling, free drawing, etc.)

Adaptivity (capacity to learn, to adjust, etc.)

The dimension of *adaptivity* is meant to be the "general dimension," a need for which is cited by several theorists. If it is a general dimension, it should be manifest in all the specific operations cited as dimensions and it is thought to be so. Almost by definition, and certainly in terms of test scores, children's growth in intelligence, their mental maturing, means that they *change* in *desirable* directions re recall, discrimination, and the rest. Kounin, in his investigation of rigidity-flexibility, distinguishes between a feeble-minded adult and a normal child of the same *mental age* on the grounds of flexibility (28). His "flexibility" seems synonymous with our "adaptivity." Frank N. Freeman, among others, has argued that a measure of capacity to learn would be the best measure of intelligence, thus inferring that learning ability (or adaptivity) is a dimension of the general intelligence factor (20).

Moreover, a child's success on an intelligence test means that he *has* learned certain things (names, number combinations, definitions, etc.) or *can*

⁴ In this discussion of dimensions, little attention is given to the dimensions which test publishers *say they measure* by their tests because the bulk of intelligence tests (including the Binet and the Wechsler) do not represent systematic efforts to measure clearly defined dimensions. They have been designed seemingly to differentiate between people of presumably different intelligence and items have been included that are known to differentiate. Obviously the authors of such tests have had a concept of intelligence in mind, but largely they have avoided a detailed specification of dimensions except as the labels for classes of items may be called dimensions, such as verbal or nonverbal, analogies, vocabulary, block design, incongruities, digit span, etc.

learn something very quickly (to complete syllogisms, to solve thought problems, to figure out mazes, and the like). Only the subtests of rote memory (digit and sentence recall) could be said not to involve adaptivity, yet even with them the individual must *change* from no set response to a given set response. Stoddard establishes "adaptiveness to a goal" as a dimension of intelligence and by it seems to mean "adaptivity" as we define it. Though he lists it as parallel to his other dimensions, making it no more general than the rest, his listing is from such a point of view as to make all his factors more or less all pervasive. Finally, adaptivity seems to be entirely consistent with the correlations found between tests of such separate factors as spatial relations and language and it is widely used in technical and vernacular definitions of intelligence.

Indexes of Intelligence

If there is one thing most commonly "known" about intelligence, it is that IQ (of 100 or 110 or 90, etc.) is the index of intelligence. So well is this "known" that many persons find it difficult to believe that tests that do not yield an IQ (the ACE and the AGCT, for example (Appendix B)) do measure the same sort of intelligence as those that do yield an IQ. Many other persons call an intelligence test an "IQ" test. And, finally, the question "What's his IQ?" is as commonly asked about a pupil as is the question "What's his intelligence?"

Now as a matter of fact the IQ is only one of several forms in which measures of intelligence are expressed. In recent years classificatory-descriptive forms have been disregarded but a number of different indexes of rank and of scale position are in use in addition to the IQ.

MENTAL AGE

For children, mental age, abbreviated MA, is perhaps the most prevalent index of intelligence rank. The standard individual intelligence test for children, the Stanford-Binet (Appendix B), yields an MA as do nearly all the group tests in common school use. MA means simply that a given child's performance on a test is like the *average* performance on the *same* test of children of a given *chronological age*. Thus, on the California Test of Mental Maturity, Elementary 1950 S-Form (Appendix B), the average score of children who are just nine years old is 54. Suppose that Harry scores 54 on the test; then Harry's MA is nine years, no months. Chronologically, Harry may be eight years old or ten years, two months; but his Mental Age (MA) is 9-0 *because his test score is like the average test score of children who are 9-0*.

We need to understand very clearly that MA is an index of *rank* and *not* of *scale position*. MA refers to a population of children each of average intelligence ranked in ascending order of chronological age. It tells no more than where to place the child tested in this ranked population.

INTELLIGENCE QUOTIENT

The IQ, on the other hand, often does denote scale position. Except for certain adult and preschool tests, intelligence tests invariably are designed to yield an IQ or Intelligence Quotient. In its inception, IQ meant a child's Mental Age divided by his Chronological Age multiplied by 100, or $IQ = MA/CA \times 100$. It still means this for the 1937 Revision of the Binet, the "standard" individual test for children, and many group tests. Consequently, you are advised to remember the simple formula $IQ = MA/CA \times 100$. The manuals of tests that yield IQ's may be read directly and thus no computation is necessary when testing. However, pupils' records may contain an IQ or MA notation only and you may need to have the other index.

As a scale expression, an IQ always has 100 as its reference point. The 100 signifies the average for any age group and IQ's above or below 100 indicate intelligence above or below the average intelligence for the age in question. Since intelligence, as measured, seems to be distributed in a "normal" fashion (see pages 163-169 for a discussion of the normal probability curve), IQ increments may be considered fixed fractions of a Standard Deviation and thus they approximate the characteristics of scale units. In the standardization population of the 1937 Revision of the Binet, the size of the Standard Deviation was 16 IQ points. Group tests habitually are "normalized" to an IQ distribution with an SD of approximately 16.

Deviation IQ's. Some IQ's, most notably those derived from the Wechsler-Bellevue, do not mean a ratio of MA to CA but rather a deviation from a mean of 100. Because in such tests the Standard Deviation of IQ's approximates 16 (it is 15 in the Wechsler) the "deviation" IQ and the "ratio" IQ may be used interchangeably for school-age children. For adults, the ratio IQ is inapplicable except by the Stanford-Binet convention that all adults have a chronological age of 16.

The Standard Deviation unit is used directly as the basis for scale scores yielded by one important intelligence test, the AGCT (Army General Classification Test, Appendix B). The many of us who were processed by the Army during World War II were used to AGCT's of 90, 115, 105, and the like. Since by test design 100 was the average score of all "literate" trainees, 90 meant below average and 115 and 105 above. The Standard Deviation of AGCT scores is 20. Thus IQ's and AGCT's are not directly comparable even though they both are scale scores with means of 100.

PERCENTILE RANKS

A third basic expression for measure of intelligence is percentile rank. This occurs most notably in the ACE (American Council on Education Psychological Examination, Appendix B), a college-level test of academic ability. The publishers of this test have asserted that IQ's are inappropriate at the adult level and that an indication of rank within a known population is a more

defensible index of intelligence for adults. Now, because in normal distributions there is a known relationship between percentile rank and standard deviation distance, percentile indications of intelligence and IQ indications of intelligence may be related if the IQ and the percentile both refer to the same population. (See page 499 for a Table showing this relationship.) *If they do not refer to the same population, they may not be related.* On the basis of the relationship, a number of school-age intelligence tests include grade percentile equivalents of their IQ's. This facilitates studies of pupils where their achievement in subjects may be expressed in percentiles. An example of improper comparisons would be between IQ and percentile rank on a test standardized on college students. Ordinarily a percentile of 50 would equal an IQ of 100. In this case, however, a percentile of 50 would more likely equal an IQ of 115 to 120 and a percentile rank of 1 or 5 would more nearly compare with an IQ of 100.

UNRELIABILITY OF INDEXES OF INTELLIGENCE

In conclusion, it should be noted that none of these measures of intelligence is exact, as your age is exact or your weight today. Each is derived from a test that has some degree of unreliability. Because of test unreliability each raw test score is unreliable and then, of course, so is the MA, IQ, or percentile indicated. The unreliability of a test score is called its Standard Error (see page 169 for discussion of Standard Error), and the Standard Error of intelligence indexes are appreciable. One of the most reliable IQ's, that derived from the individually administered Stanford-Binet, has a Standard Error of approximately 4 in the middle of the distribution of IQ's (41). It becomes greater for higher IQ's and less for smaller ones. Thus, if a child whose Binet IQ was found to be 104 were retested a large number of times, the probability would be that 68 per cent of the IQ's found for him would vary from 100 to 108, 96 per cent from 96 to 112, 99 per cent from 92 to 116, and so on. But with no certainty could you say at any time that his IQ was 104 exactly. Since other intelligence tests are hardly more reliable than the Binet and many are known to be less, it is safe to assume a Standard Error of score of the same or greater size for IQ's derived from them.

The General Nature of Intelligence Tests

The construction of intelligence tests is not a problem for the teacher or school administrator. There are several score published tests that have sufficient reliability for school use and that, collectively, are pertinent for all ages of children, youth, and adults and for the special needs of different types of schools. In this section we shall not attempt to catalog these tests nor to describe any particular specimens. Annotated listings of intelligence tests occur in Appendix B, pages 475-479, along with listings of other standardized tests. Rather we shall examine their common characteristics so that you may have a more reasonable basis for selecting a test and for interpreting its scores.

STRUCTURE AND ITEMS

Essentially the great majority of intelligence tests measure the performance of an individual relative to things important for achievement in school or for success in other abstract verbal activities. The performances may require recall of a specified learned thing, as,

$$3 \times 17 = \quad (a) \ 20 \quad (b) \ 50 \quad (c) \ 51 \quad (d) \ 41$$

or

Aunt (means) (1) grandmother (2) mother's sister
(3) a neighbor woman (4) the same
as sister.

The performances may involve instead the recall and application of some general learnings, as,

Rearrange these words to form a sentence:

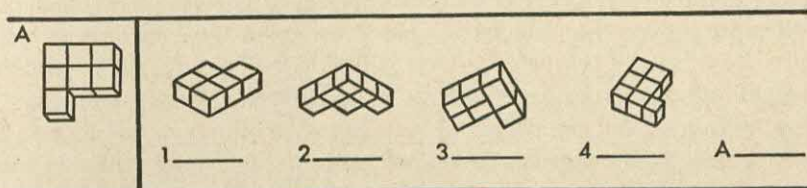
left eat dogs dinner are
bones that from

Or the performances may not depend upon any prior learning but simply require a given adaptive behavior, as,

Repeat these digits just as you hear them. 4 3 9 6 2

or

Find a drawing that is a different view of the object in the first drawing.



(From *California Test of Mental Maturity—Intermediate Series, 1946 Revision*. Copyright 1936–1951, California Test Bureau.)

By far the largest number of intelligence test items are of the learned performance variety. This necessitates the assumption on the part of test-makers that all who take the test have had equivalent opportunity to learn the required things. This fact of intelligence tests makes them most appropriate for school-age children who have had consecutive experience in public schools. It tends to diminish their validity for some ethnic and socioeconomic groups, a factor that we shall explore in a moment.

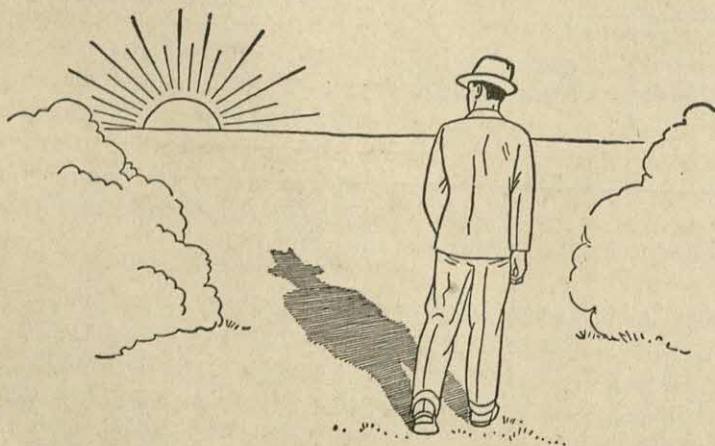
Nonverbal Items. Because verbal items require reading (or good listening vocabulary in the case of spoken questions in primary tests and in in-

dividually administered tests in some instances) pupils who read well do better than those who read less well on verbal items. For this reason, most tests applicable to the elementary grades and some appropriate for the secondary grades include some nonverbal items. Many have a sufficient number of nonverbal items to warrant their categorization and separate scoring as such. Generally speaking, tests devised for older children contain fewer nonverbal items. There are certain tests that presume to be strictly nonverbal (see Appendix, page 477). It is necessary to remember, in the case of nonverbal items or tests, that the directions usually still are verbal and that the responses often must be verbal. Two examples of nonverbal, or as they are sometimes called, performance items are:

1. The usual "What part is missing" item, occurring in many primary level tests.



2. The less usual incongruous element item. In this case a picture is shown that contains something awry and the testee is asked to indicate the incongruous element. The example shown is from the 1937 Revision of the Stanford-Binet.



© 1937, 8

(From *Revised Stanford-Binet Intelligence Scale*, L. M. Terman and M. A. Merrill, Houghton Mifflin.)

Group Tests. Intelligence tests frequently are classified as group and individual tests. The former are, without exception, guided response instruments and the majority of their items are of the multiple-choice variety. Nearly all of them are administrable in two hours or less, with many requiring no more than forty minutes. Typically the younger the age for which the test is pegged, the shorter the test.⁵

Nearly all group tests, except those for primary grades, contain some arithmetic problems. In an examination of 22 high school and adult tests, Kenney found a range from 1 to 56 per cent of the items being arithmetic (27). Naming synonyms, completing series (2-6-18- ?) and answering analogies questions are further typical item types.

Group tests, as the classification implies, are administered to many children at once and may for this reason have less validity and reliability than do individual tests. It is intended that group tests be administered by teachers without special training. The instructions in accompanying manuals and on the tests usually are clear and comprehensive.

Individual Tests. Individual intelligence tests are fewer in number and, as a rule, require special training and skill for administration. The most popular for school-age children and youth, the Stanford-Binet and now the WISC or Wechsler Intelligence Scale for Children, are essentially structured interviews with predetermined tasks and predetermined bases for recording and rating responses. Some of the items entail guided responses, but the greater number allow the testee to respond as he will. These free responses require that the examiner use his judgment as well as a key. For example, in the Binet, one test asks that the child look at a picture and then "tell you about it." From his telling you must gauge what credit to allow him. To help you and to standardize scoring, the test manual contains a detailed analysis schedule for responses.

Individual tests are considered more valid and more reliable and hence are used to verify group test findings, to detect mental deficiency and to diagnose intellectual strengths and weaknesses. Psychometrists and clinical psychologists usually learn to administer individual tests as part of their training and perform this service for schools.

PROFILES AND DIAGNOSIS

In recent years increased attention has been given to differential aspects of intelligence. Frequently it has been said with truth that an IQ hides as much as it shows. A child might do poorly on number items, well on definitions, and average on digit span to get his IQ of 100. Another might do well on number items, poorly on digit span, and average on definitions to get the same IQ of 100. Now that some relatively distinct factors (or, as we would name them, dimensions) of intelligence have been determined statistically and by reason-

⁵ Administration times for specific tests are in Appendix B.

ing, some test designers have attempted to measure these factors separately and to furnish separate scores on each such factor. These scores are often recorded on profile sheets comparable to the one shown in Figure 51.

The factor scores and/or their profile representations furnish information which single MA's, IQ's, percentiles, or Standard Scores do not. They tell such things as whether a given IQ represents much memory but little adaptivity, or little memory and much adaptivity. They may suggest the extent to which test nervousness or anxiety has depressed performance since this condition affects exact recall and attention-demanding performances the most. And thus the factor scores or profiles have some diagnostic significance. To know that a youngster of Greek parentage scores very low on language factors but fairly high on number factors is to know that probably he is very bright but needs tutoring in language. To know that a child with good grades in arithmetic scores low on reasoning but high on memory is to be able to predict that he will likely have more success in business arithmetic than in algebra.

In using profiles or factor scores, it should be recognized that the part measures are less reliable than the total score and that group tests *are not clinical instruments*. But for detailed information, for leads to further measurement, and for referral purposes they are of far greater value than single indexes.

RELIABILITY AND VALIDITY OF INTELLIGENCE TESTS

Such precise data as are available for the reliability and validity of specific tests are presented in the Appendix along with other brief identifying notes on the tests. Here it will be enough for us to consider the matter of reliability and validity in general.

Using split-half and some comparable-form or test-retest coefficients of correlation as indexes of reliability, the bulk of published tests for school-age children have reliabilities comparable to or in excess of the reliabilities of standardized tests generally. Coefficients of reliability tend to range from .88 to .97. Thus, if any standardized tests can be considered reliable enough for school use, intelligence tests can. Preschool tests and, in particular, infant scales are much less reliable. Except for extremely retarded or extremely accelerated infants, IQ's should not, in the opinion of the authors, be predicted on the basis of these tests. Somewhat more confidence may be placed in measurements of three-, four-, and five-year-olds; but even here IQ's should be asserted with great reservation.

The validity of intelligence tests has been disputed in recent years. Without detailing the arguments or without examining the great body of research that bears on their validity, we may note the bases for assertions of validity for tests and two points at which the validity of tests seems to be vulnerable.

Bases for Claims of Validity. Tests commonly base their claims of validity on correlations with other tests, on correlations with school marks and other "external" indicators of intelligence, on analyses of growth curves, and

1946 REVISION

Intermediate
Grades 7-10
and Adults

CALIFORNIA TEST OF MENTAL MATURITY—INTERMEDIATE SERIES

Devised by Elizabeth T. Sullivan, Willis W. Clark, and Ernest W. Tieg

Name..... Occupation or Grade.....

Date..... Age..... Last Birthday..... Sex: M-F

Teacher or
Examiner..... School or
Organization.....

TEST	FACTOR	Possible Score	Score	0	28	29	40
1.	Visual Acuity	40		Low		Average	High
2.	Auditory Acuity	15		Low	9	Average	15
3.	Motor Co-ordination	20		Low	10	18	High

DIAGNOSTIC PROFILE
(Chart scores here)

TEST	FACTOR	Possible Score	Score	Mo.	96	120	132	144	156	168	180	192	204	216	240	288	Per- cent- ile Rank
				Yr.	8.0	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	20.0	24.0	
A.	Memory	50			5	20	25	30		35	40	45	47	49			
	4. Immediate Recall°	30			4	12	13	14	15	16	17	18	19	20	21	22	23
	5. Delayed Recall (p.16)	20			1	5	10	12	14	15	16	17	18	19	20	21	22
B.	Spatial Relationships	45			8	15	20	22	25	27	30	32	34	36	39	43	
	6. Sensing Right and Left°	20			5	10	13	14	15	16	17	18	19	20	21	22	23
	7. Manipulation of Areas°	15			2	5	6	7	8	9	10	11	12	13	14	15	16
	8. Foresight in Spatial Sit'ns°	10			1	2	3	4	5	6	7	8	9	10	11	12	13
C.	Logical Reasoning	60			4	10	15	20	25	30	35	40	45	50	57		
	9. Opposites°	15			2	5	6	7	8	9	10	11	12	13	14	15	16
	10. Similarities°	15			1	3	4	5	6	7	8	9	10	11	12	13	14
	11. Analogies°	15			1	2	3	4	5	6	7	8	9	10	11	12	13
	12. Inference (p. 14)	15			0	1	2	3	4	5	6	7	8	9	10	11	12
D.	Numerical Reasoning	45			1	5	10	15	20	25	30	35	40	45	50	57	
	12. Number Series°	15			0	1	2	3	4	5	6	7	8	9	10	11	12
	13. Numerical Quantity°	15			0	1	2	3	4	5	6	7	8	9	10	11	12
	14. Numerical Quantity	15			0	1	2	3	4	5	6	7	8	9	10	11	12
E.	16. Vocabulary	50			1	5	10	15	20	25	30	35	40	45	50	57	
Total	Mental Factors	250			44	60	74	80	90	100	110	120	130	140	150	160	170
	(A+B+C+D+E)																
F.	Language Factors	100			15	20	25	30	35	40	45	50	55	60	65	70	75
	(5+14+15+16)																
G.	Non-Language Factors	150			30	40	50	60	70	80	90	100	110	120	130	140	150
	(Total Mental Factors Minus F)																
Chronological	Age				96	120	132	144	156	168	180	192	204	216	240	288	
Actual	Grade Placement				3	4	5	6	7	8	9	10	11	12	13	14	15
	(Grade pupil is in.)																

*Non-Language Tests

SUMMARY OF DATA

Total Mental Factors
 F. Language Factors
 G. Non-Language Factors

Copyright, 1934, 1937, 1943, 1946, by California Test Bureau
 Published by California Test Bureau
 5916 Hollywood Boulevard, Los Angeles 28, California

*Age 16 and older, divide by 192 months.

Figure 51. Profile Page of California Test of Mental Maturity—Intermediate Series, 1946 Revision. Copyright 1936–1951, California Test Bureau.

on internal evidence. The first basis involves giving two intelligence tests to the same group and obtaining correlations between the scores. The Stanford-Binet is frequently the second test. Sometimes a battery of tests is used instead of just a single criterion. Such a process is good, of course, and high correlations mean high validities as long as we may *assume the validity* of the criterion test. The second basis is that of comparing IQ's with school marks or teachers' ratings. The correlations between test scores and these variables generally are lower than between tests. Yet this basis does involve an external criterion that seemingly should be related to intelligence. In the third place, test-makers examine the scores of successively older children and, when a smooth rise in mean scores is found, they say this indicates validity. The reasoning involved here is that intelligence "grows" smoothly and if test scores also "grow" smoothly they must be measuring intelligence. Finally, statisticians analyze tabulations and graphs of the scores of tests given to persons of known characteristics and from this they infer validity or the lack of it. If, for example, the graph of a large unselected population has a "normal" shape, validity is indicated since intelligence is presumed to be "normally" distributed. Moreover, validity is to some extent a matter of inspection. The test says it measures memory. Look at the test and see what items there are that require recall for correct answering.

Verbal Bias. With these ways of appraising validity in mind, let us examine two aspects of intelligence tests often criticized when validity is discussed. In the first place, the written tests require from a little to a great deal of reading and the oral tests require enough oral vocabulary to understand directions and to say the answers. Unquestionably, those children who have the greater reading and speaking skill score higher on the intelligence tests usually employed in schools, and those who have the less verbal skill score lower because of the difference in verbal skill. Time and again this point has been probed by researchers and always the generalization is verified. One recent example of such research is Darcy's (13) in which he found that 235 children of Puerto Rican parentage had mean IQ's of 79.6 on the Pintner Verbal Test but mean IQ's of 87.8 on the Pintner Non-Language Test. Another example is the comparison made by Bond and Fay (8) of the performance on individual items of the Stanford-Binet scale of good and poor readers. When children were equated as to mental age, the investigators found that good readers performed significantly better on knowledge and word use items; the poor readers better on nonverbal and memory items. They conclude that poor readers are underrated and good readers are *overrated* on the Binet.

One of the authors has administered Binets to many school children of Mexican parents who spoke their native tongue in the home. It is his impression that two out of three of these children had been found with low IQ's and had been considered for mentally retarded classes *not for any want of brightness but for want of opportunity to learn the English language*. In particular does he remember one boy of eight or so who was unable to identify any of

the pictures of common objects or to define any of the words in the vocabulary list yet he *seemed to know* many of them. Finally, the examiner asked him to name objects and to define words in Spanish, which he did at a great rate. But the examiner didn't speak Spanish so a teacher had to be found to translate. With her help the boy named and defined to a degree commensurate with his age.

Now it may be contended, with justice, that this verbal bias of intelligence tests does not necessarily invalidate them. Speaking and reading, writing and listening, of all human behaviors, perhaps necessitate the most discrimination, memory, generalization, etc., and hence you should expect the children more skilled verbally to be the more intelligent. Moreover, the primary use of intelligence tests is for educational prediction and placement and, assuredly, the common curriculum is a verbal one. So when the children tested have had seemingly normal opportunity for learning the English language or when their scores on nonverbal elements are equivalent to those on verbal elements, an IQ derived from ordinary tests is meaningful. But, when these conditions do not obtain, we advise that you view such IQ's with suspicion if by IQ you mean a measure of intelligence as defined here.

Culture Bias. Intimately related to the language bias of tests is what is now called their culture bias. Davis, Havighurst, Eells, and others (16, 17, 19, 25), principally of the University of Chicago, have for many years been contending that intelligence tests discriminate against children of lower class socioeconomic status. Essentially, they argue that American published intelligence tests have middle-class content. Lower class pupils have little or no access to this content and thus have restricted opportunity to learn what they must have learned if they are to succeed on the tests. To examine this possibility of cultural bias, let us consider the following items selected from a widely used group intelligence test.⁶

In three sections of the test, proper answering requires the proper identification of small pictures portraying many different objects, persons, and activities of American life. Among these are such things as a chandelier, a meat grinder, skis, a knapsack, a radio, an Egyptian statuette, a wolf and a lamb, a dripolator and percolator, a western saddle, ice tongs, and, it is necessary to confess, a number of objects with which the authors have had no previous experience and hence which they cannot identify. In the vocabulary section occur *entreat*, *facetious*, *disparage*, and *presage* to say nothing of *vertigo* and *quondam*.

So far as the objects are concerned it is open to question whether children twelve to fifteen in age, the group for which the test is designed, raised in migratory labor camps in California, in a tenement in Chicago, or on a rented farm in Arkansas would have had direct or even indirect experience with all these objects. As to the words, it is thought that only families of considerable

⁶ The test is thought to have neither more nor less cultural bias than tests in general.

education and having interest in serious reading would ever use the words or even own books containing them. Yet, the words represent ideas of no great complexity. Lower class and even many middle-class people simply use different expressions for them, i.e.,

for entreat: beg

“ facetious: funny

“ disparage: belittle

for presage: predict

“ vertigo: dizzy

“ quondam: former

It seems evident to us that intelligence tests do have a cultural bias and that the bias is toward that of the culture of teachers. As predictors of success in school, however, this should not invalidate the tests. For, as we know, the culture of teachers largely constitutes the culture of the schools and to be or to become proficient in this culture is to make high marks. The fact of bias does necessitate care in interpreting test scores. It requires that decisions about children of lower class environments be made only after a careful analysis of test performance and, if needed, retesting. Some tests, and portions of most tests, are relatively culture-free. Davis and Eells have published one, which they say has scarcely any cultural bias (15).

IQ VARIABILITY AS AMONG TESTS

We have mentioned that any intelligence test has some degree of unreliability and hence, on repeated application of the same test, it should be anticipated that a pupil's IQ will vary by several IQ points. Here we need to observe that *different tests* administered to the same pupil are likely to yield different IQ's. Test designers use different standardization populations. The shapes of the distributions of scores vary from test to test. The standard deviations, basic units of variability for distributions of scores, are greater for some tests than others. Each test has a different proportion of mathematics items, vocabulary items, nonverbal items, etc. For these and perhaps other reasons it has been impossible for test designers to produce tests that are mutually comparable for all their range despite the serious efforts of most test designers to do just that.

As a rule, IQ's from one test are most comparable to IQ's from another test at the middle of the distribution or, let us say, from IQ 90 to IQ 110. As extremely low or extremely high IQ's are encountered, the IQ's often become somewhat dissimilar. The size of these variations is exemplified by Lennon's study of three tests applied to 1,200 high school students (29). He compared the Terman-McNemar Test of Mental Ability with the Otis Quick Scoring Mental Ability Tests and the Pintner General Ability Tests. Among his findings were that Pintner IQ's were 2 to 5 points lower than Terman-McNemar's and that the Otis had a smaller SD than did the Terman-McNemar. For this reason an IQ of 66 on the Terman-McNemar meant an IQ of 76 on the Otis, and an IQ of 131 on the Otis meant one of 141 on the Terman-McNemar.

Whether a given test's IQ tends to run high or low, and how much, is often noted in critical reviews of tests. These occur in many professional journals (*Educational and Psychological Measurement*, for one) and most notably in Buros' *Mental Measurement Yearbooks*. In addition, the school's or district's psychometrist and/or psychologist should be able to give advice about the IQ's of different tests. Where judgments about mental development over a period of time are to be made, the variability among different tests is particularly critical. If possible, these judgments should be based on readministrations of the same test or its comparable forms.

SOURCES OF INFORMATION ABOUT INTELLIGENCE TESTS

A number of intelligence tests and several publishers of them are cited in Appendix B, pages 475–479. Additional bibliographic sources of information are Buros' *Mental Measurements Yearbook* (9), the *Encyclopedia of Educational Research* (30), professional journals (psychological as well as educational), and publishers' catalogues. The *Yearbook* generally is considered the source. Certainly, in it tests are most likely to be viewed critically and it is the only source intending to be exhaustive. Publishers' catalogues provide the most complete statements of price, length, scoring, etc. Some provide reliability indexes but they tend to stress the wide applicability of any test rather than its limitations. The manual of a given test usually describes the way in which it was standardized and provides evidence of reliability and validity.

Administration and Scoring of Group Intelligence Tests

Practices in the administration and scoring of group intelligence tests vary from school district to school district. In some, a special person, usually designated as a psychometrist or psychologist, administers and scores group tests as well as individual tests. In others regular teachers (or in high schools, counselors or orientation teachers) administer and score the group tests. In a few cases a teacher administers and a specialist scores or vice versa. Districts having access to machine scoring devices typically use machine answer sheets, where usable, and the tests are scored by a clerk or technician.

Precise directions for administering and scoring tests are contained in test manuals. The general principles of test administration were presented in Chapter 6, pages 118–121, and are applicable to intelligence tests. When and if you administer an intelligence test, you are asked to be extremely attentive to the directions and to adhere closely to principles of good test administration. No other single test result is likely to have such an important bearing on how a child will be handled in school.

Good test administration, you may recall, is more than a mechanical matter. It involves the way children feel about the test situation and what you do to make them feel right. As you know, the technical term for this feeling matter is rapport. Unless rapport is adequate, don't administer the test. If Betty seems frightened and you can't coax her out of it, have her take

it another day. If children appear to rush because a passing bell is coming soon *and it isn't a timed test*, invalidate that portion of the test and retest them on another day if necessary. Should unusual distractions occur during the testing period, make note of them and judge test results in the light of them. You may wish to discount them. In short, you are enjoined to administer an intelligence test *only* in the best manner and *only* under optimum conditions.

Evaluation of Intelligence

Few evaluative standards exist for intelligence and, for the most part, evaluation is not involved in appraisals of intelligence. Since intelligence is presumed to be a factor that bears on school achievement and not a matter of achievement, teachers do not mark or grade a child in it as they do in arithmetic or public speaking. When a school psychologist measures a pupil's MA as part of his effort to determine how best to handle him, fact finding, not judgment, is involved. Judgment or evaluation, as we define it, entered when a teacher referred the pupil because of *misbehavior* or *poor* achievement or presumed *neurotic* symptoms. And later the student's adjustment will be evaluated to determine if he should return to this teacher's class. But the fact of a low or a high MA is simply a diagnostic reality, not a matter for judgment.

Mental Deficiency. The evaluations expressed regarding intelligence and what standards there are usually have to do with mental deficiency, genius, and academic or vocational placement. Many states have laws and many school districts have regulations as to what constitutes mental deficiency. The aspects of these laws or regulations that concern teachers are standards of educability. Typically, these are very general statements about the child's ability to profit from instruction, to learn a trade, to be self-reliant, etc. After individual testing the examiner judges whether or not the child should be kept in the regular school situation or given the special treatment provided for those who fall below the legal standard. By custom, perhaps also for good reason, an IQ of 65-70 tends to be the rule-of-thumb line between those judged mentally deficient and those judged normal but dull. Standards (in terms of IQ's) for the further extremities of mental deficiency, those that require institutional confinement or some degree of permanent custodianship, are of little significance to regular teachers since children so extremely deficient rarely reach the schools.

Gifted Children. At the other end of the intelligence spectrum lie those who are called "gifted" or "precocious" or "geniuses." In recent years much less effort than heretofore has been given to precise classifications of genius. Genius, as you are aware, is simply a word we use for those whose achievement we greatly admire. It has no precise signification unless we wish to give it one. Terman, in his studies of "genius," (42), used a specified IQ, 140, as a basis for admission to his group of gifted children. However, lower IQ levels have been designated for genius by other investigators, 130, 125, even 120; and, on the other hand, it has been asserted that the term should be reserved

for those of IQ 180 and above (30:505). Thus, if in a school you wish to find your "gifted" children, you will have also to find a standard for them.

Subject Placement. Vague as intelligence standards are for mental deficiency, they are even more vague for subject and vocational placement. From many studies (30:878-883) a positive correlation has been established between intelligence and achievement. Yet the precise degree of mentality (expressed as an IQ, percentile, etc.) essential to success in any subject or vocation is largely undetermined and it may be undeterminable. It is customary to restrict study of higher mathematics, laboratory science, and foreign language in the high school to those of "better than average" intelligence, but neither research nor practice warrants designation of precise minimum MA or IQ levels to these or any other subjects. In general, it may be said only that as subjects are more verbal and abstract, a higher level of intelligence is required, and, as they are less verbal and abstract, children of less intelligence have more chance of success.

Vocational Placement. Similarly, in vocational or professional training, positive correlations have been found between success and intelligence in many areas (30:886-890). No minimum levels of intelligence have been established and we may be sure only of the general principle that the more complex, abstract, and self-directed vocations require the more intelligence. In applying this principle, schools and industries frequently adopt exact intelligence levels as cutting points in selecting trainees. They do this arbitrarily and must defend their designated levels on the basis of "it works for us."

Measures of Intelligence and Educational Practice

As we have studied the several aspects of measuring intelligence—dimensions, forms, procedures and, just now, standards—we necessarily have suggested many applications of intelligence measures to educational practice. Now, to conclude the chapter, we wish to state some cautions to be observed in the use of IQ's and MA's and then to specify certain school uses for measures of intelligence.

SOME CAUTIONS ABOUT THE USE OF IQ'S AND MA'S

Heredity and Environment. To begin, the intelligence measured by a test is a function of both heredity and environment. A child's parentage and ancestry *does*, through genetic transmission, affect his IQ. But his preschool and school experiences also *do* affect his IQ. Thus a social worker may not with justice declare that Bill's case is hopeless because he is born to illiterate and dull parents. On the other hand, he may not expect to change Bill's intelligence radically by placing him in a good foster home. Special investigations of twins raised in different homes, of foster children-foster parent relationship, of racial and national differences in intelligence, and of famous and infamous families all indicate that environment and heredity contribute to intelligence in unknown degrees (23, 31, 32, 36, 37).

IQ Variability. Secondly, it may be observed that the *IQ is not fixed*. We have seen that any test's inherent unreliability makes it likely that retesting will produce different scores. We have observed, moreover, that the IQ's from different tests are likely to vary because of differences in items and in shapes of standardization distributions. In addition, IQ's are not fixed because test performance is a function of the pupil's health and attitude as well as his intelligence. As these improve or deteriorate so does his test score. They are not fixed because specific instruction on the content and method of the tests (not cribbing, of course) may improve test performance (5). And they are not fixed because intelligence itself may "grow" at a different rate from that presumed by the test designer and hence the IQ's may vary. It is impossible to predict just how much IQ variation is likely to be encountered although researchers have tried to find out (18). As a rough guide, we suggest that you consider variation of up to five points as usual, up to eight points as nothing to get excited about, but over eight points as a matter that deserves specific explanation.

Designed for School Children. A third point of caution is that intelligence quotients and mental ages are designed for school-age children who are attending school and should be so used. We have seen (page 372), that pre-school and particularly infant tests yield IQ's that are inefficient predictors of school age IQ's. For adults the concepts of mental age and of IQ as a ratio are clearly inappropriate. And intelligence tests presume a common continuous experience on the part of those tested. Children not in school are those least likely to have this common continuous experience.

Relation to School Marks. Then, you have heard many times that IQ's bear a correlation of about .50 to school marks. Actually, there is less relation with some subjects, the performance ones, and more with others, the predominantly verbal ones. Each study finds a different correlation figure and each test seems to show a different figure, but for the elementary and high school grades the correlations tend to cluster around .50. Now it is important to know that this .50 *does not mean* 50 per cent. It does not mean that half of the variation in marks among students is to be attributed to variations in intelligence. If it connotes any percentage, *it should connote 25 per cent*, this being the percentage of variation that a coefficient of correlation of .50 indicates is explained by the correlated variable. (See page 180 for further explanation of this.)

Parental Interpretation of an IQ. As a fifth caution, we need to be reminded that parents tend to think that an IQ means a given grade of intelligence and that is all there is to it. A relatively small percentage may understand it as the teacher understands it. The great remainder will not. These will not treat an IQ as they would an index of height or weight. Rather, they will worry about it or wish to brag about it to their neighbors. They will view their child's IQ as a proper measure of a tangible mental substance, intelligence, and the measure, whether low or high, will have great significance

for their egos. This is why we maintain the rule: *parents are not to be told their children's IQ's.*

Individual and Group Prediction. Finally, in the way of cautions, let us differentiate between the use of IQ's for group and for individual prediction. In Chapter 8 we discussed the matter of probability in behavioral measurement and we saw, among other things, that the precision of group measures was related to the size of groups measured. Thus an average for a group of 10 is less precise than an average for a group of 100. If you will consider that one pupil is the smallest possible group, then you can see readily that the IQ found for one pupil is more of an approximation than is the average IQ found for 100 pupils. Therefore, since precision in measurement affects accuracy of prediction, we must consider it far safer to make predictions about groups, having measured their intelligence, than to predict about individuals. For example, you may say with assurance that 100 children having IQ's of 110 or more will as a group do better in algebra than another 100 having IQ's of 109 and less. But you should have much less confidence that John with an IQ of 115 will make a higher grade in algebra than Tom with an IQ of 105. The odds are that he will, but it's far from a safe bet.

THE USES OF INTELLIGENCE TESTS

IQ's and MA's, intelligence percentiles, and standard scores are useful in many ways to teachers and school administrators. They should be used only with full understanding of their significance and fallibility and with such cautions as we have described. However, they should be used. Simply to administer an intelligence test or to know a pupil's score accomplishes nothing in itself. What matters are the things that may be done on the basis of intelligence testing that could not be done as efficiently without it. Among the ways in which the scores of intelligence tests may be used are the following:

- (a) In determining reading readiness.
- (b) In determining whether or not any pupil's achievement approximates his potential.
- (c) In determining eligibility for subjects or courses requiring a high degree of intelligence.
- (d) In establishing sections of a grade or a course differentiated according to ability
- (e) In vocational and educational guidance in secondary schools.
- (f) In selection of reading materials for given classes.
- (g) In assigning pupils to special classes or curriculums.

Summary

Intelligence is a construct to explain differences among individuals as to intellectual aspects of behavior. Its dimensions vary from theorist to theorist and from test to test. In this context the dimensions are considered to be

recall, discrimination, symbolization, abstraction, invention, and adaptivity. Differences in intelligence are expressed primarily by Intelligence Quotients (IQ) and Mental Ages (MA), but Percentile Rank and Standard Scores are in use as well. The last is the measure preferred by the Armed Services and percentile rank is employed by the most widely used college-level intelligence test.

Numerous tests administrable to groups are available for every school grade and for adults. The Revised Stanford-Binet Scale and the Wechsler-Bellevue Scales are the most prominent individual tests. Tests for preschool children necessarily require individual administration and are less reliable than school-age measures. Precise evaluative standards do not exist for intelligence. Typically, evaluation is undertaken only in determining mental deficiency and genius and in academic and vocational selection.

Measurements of intelligence are invaluable for educational practice, but their use necessitates observance of certain cautions. The IQ reflects both heredity and experience and is not a fixed quantity. It is reasonably constant and the more so during the elementary grades. Since the IQ and MA indexes were designed for school-age children and for children attending school, they should be so used. The correlation between school marks and IQ is about .50, higher r 's being found for abstract verbal subjects and lower r 's for performance and activity subjects. Parents will tend to regard the IQ with exaggerated significance. Finally, predictions based on intelligence testing are more accurate for groups than for individuals.

EXERCISES

1. Inspect a group intelligence test and indicate the following:
 - a. Items that have a cultural bias.
 - b. Items that place a premium on language skill.
 - c. The dimensions of intelligence that the test seems to measure.
2. Study the manuals of several tests and make a written comparison of the tests according to:
 - a. Clarity of directions.
 - b. Adequacy of standardization.
 - c. Reliability.
 - d. Coverage of the dimensions of intelligence discussed in this chapter (see page 365).
3. Compute the missing index in each of the following. Consider that they relate to the Stanford Revision of the Binet.
 - a. CA = 12, MA = 11, IQ = ?
 - b. CA = 9, IQ = 110, MA = ?
 - c. CA = 18, MA = 18-6, IQ = ?

4. Define the important differences in administration and use between a group test of intelligence and an individual one.
5. Administer a group test to a class, score the papers, and compute Mental Ages and IQ's for each pupil tested.
6. Explain the difference between an IQ and a Percentile Rank as indexes of the intelligence of adults. Between Mental Age and Percentile Rank as applied to school age children.
7. Discuss the different educational uses of measures of intelligence, giving attention to the validity and reliability of intelligence tests.

BIBLIOGRAPHY

1. Allen, R. M., and Besell, H., "Intercorrelations Among Group Verbal and Non-Verbal Tests of Intelligence," *Journal of Educational Research*, 43:394 f., January, 1950.
2. Anastasi, Anne, *Psychological Testing*. New York: The Macmillan Co., 1954, chaps. 8-12, 14.
3. Anderson, John E., "The Limitations of Infant and Pre-School Tests in the Measurement of Intelligence," *Journal of Psychology*, 1939, 8:351-379.
4. Binet, Alfred, "Le Développement de l'Intelligence chez les Enfants," *L'Année Psychologique*, 14:1-94, 1908.
5. Boger, J. H., "Experimental Study of the Effects of Perceptual Training on Group I.Q. Test Scores of Elementary Pupils in Rural Ungraded Schools," *Journal of Educational Research*, 46:43-52, September, 1952.
6. Bolton, F. B., "Value of Several Intelligence Tests for Predicting Scholastic Achievement," *Journal of Educational Research*, 41:133-138, October, 1947.
7. Britton, I. H. "Influence of Social Class upon Performance on the Draw-a-Man Test," *Journal of Educational Psychology*, 45:44-51, January, 1954.
8. Bond, Guy, and Fay, Leo, "A Comparison of the Performance of Good and Poor Readers on the Individual Items of the Stanford-Binet Scale, Forms L & M," *Journal of Educational Research*, 43:475-479, February, 1950.
9. Buros, Oscar K., *Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953.
10. Clark, W. W., "Evaluating Scholastic Achievement in Basic Skills in Relation to Mental Ability," *Journal of Educational Research*, 46:179-191, November, 1952.
11. Coleman, William, and Ward, Annie W., "A Comparison of Davis-Eells and Kuhlman-Finch Scores of Children from High and Low Socio-Economic Status," *Journal of Educational Psychology*, 46:465-469, December, 1955.
12. Cornell, E. L., and Gillette, A. L., "Construction and Educational Significance of Intelligence Tests," *Review of Educational Research*, 20:17-26, February, 1950.
13. Darcy, N. T., "Performance of Bi-Lingual Puerto Rican Children on Verbal and Non-Language Tests of Intelligence," *Journal of Educational Research*, 45:499-506, March, 1952.
14. ———, "Review of the Literature on the Effects of Bi-Lingualism Upon the Measurement of Intelligence," *Pedagogical Seminary*, 82:21-57, March, 1953.

15. Davis, Allison, and Eells, Kenneth, *Davis-Eells Games*. Yonkers: World Book Company, 1952.
16. Davis, Allison, and Hess, R., "How Fair Is an I.Q. Test?" *Science Digest*, 29:43-47, March, 1951.
17. Davis, Allison, and Havighurst, R. J., "Measurement of Mental Systems," *Science Monthly*, 66:301-316, April, 1948.
18. Eichhorn, D. H., and Jones, H. E., "Development of Mental Functions," *Review of Educational Research*, 22:421-438, December, 1952.
19. Eells, Kenneth, and others, *Intelligence and Cultural Differences*. Chicago: University of Chicago Press, 1951.
20. Freeman, Frank N., *Mental Tests*. Boston: Houghton Mifflin Co., 1926, chap. 18.
21. Freeman, F. S., *Theory and Practice of Psychological Testing*. New York: Henry Holt & Co., Inc., 1950, chaps. 3-9.
22. Gilliland, A. R., "Measurement of the Mentality of Infants," *Child Development*, 19:155-158, September, 1948.
23. ———, "Socio-Economic Status and Race as Factors in Infant Intelligence Test Scores," *Child Development*, 72:271-273, December, 1951.
24. Goodenough, F. L., *Mental Testing*. New York: Rinehart & Co., 1949.
25. Havighurst, R. J., "Culture and the I.Q.," *School Review*, 57:187-189, April, 1949.
26. Jordan, A. M., "Efficiency of Group Tests of Intelligence in Discovering the Mentally Deficient," *High School Journal*, 31:73-94, March, 1948.
27. Kenney, J. J., "Mathematics in Group Intelligence Tests," *Journal of Educational Research*, 44:129-330, October, 1950.
28. Kounin, Jacob, *Experimental Studies of Rigidity as a Function of Age and Feeble-mindedness*. Iowa City: University of Iowa Library, 1939, ms.
29. Lennon, R. L., "A Comparison of Results of Three Intelligence Tests," *Test Service Note Book*. Yonkers, N. Y.: World Book Company, 1952, No. 11, 4 pp.
30. Monroe, W. S. (ed.), *Encyclopedia of Educational Research*. New York: The Macmillan Co., 1950.
31. N.S.S.E., *Intelligence: Its Nature and Nurture*. 39th Yearbook, Parts I and II. Bloomington, Ill.: Public School Publishing Company, 1940.
32. Newman, H. H., Freeman, F. N., and Holzinger, K. J., *Twins: A Study of Heredity and Environment*. Chicago: University of Chicago Press, 1937.
33. Otis, A. S., "Measurement of Mental Systems: Reply with Rejoinder," *Science Monthly*, 67:312-314, October, 1948.
34. *Otis Self-Administering Tests of Mental Ability*, Higher Examination: Form C. Yonkers: World Book Company, 1928.
35. Ramaseshan, R. S., "Note on the Validity of the Mental Age Concept," *Journal of Educational Psychology*, 41:56-58, January, 1950.
36. Shuttleworth, F. K., "The Nature Versus Nurture Problem: II. The Contributions of Nature and Nurture to Individual Differences in Intelligence," *Journal of Educational Psychology*, 26:655-681, 1935.
37. Skodak, Marie, and Skeels, H. H., "A Follow-Up Study of Children in Adoptive Homes," *Journal of Genetic Psychology*, 66:21-58, 1945.
38. Spearman, Charles, *Abilities of Man*. New York: The Macmillan Co., 1927.

39. Stanley, J. C., Jr., "Development and Application of Tests of General Mental Ability," *Review of Educational Research*, 23:11-32, February, 1953.
40. Stoddard, George D., *The Meaning of Intelligence*. New York: The Macmillan Co., 1947, chap. 1.
41. Terman, Lewis M., and Merrill, Maud A., *Measuring Intelligence*. Boston: Houghton Mifflin Co., 1937.
42. Terman, Lewis M., and others, *Mental and Physical Traits of a Thousand Gifted Children*. Genetic Studies of Genius, Vol. I, Stanford University Press, 1925, 641 pp.
43. Thurstone, L. L., "Primary Mental Abilities," *Psychometric Monographs*, 1938, No. 1.
44. ———, "Theories of Intelligence," *Science Monthly*, 62:101-112, 1946.
45. Walter, W. Grey, *The Living Brain*. New York: W. W. Norton & Co., Inc., 1953.
46. Wechsler, David, *The Measurement of Adult Intelligence*. Baltimore: The Williams and Wilkins Company, 1944.

CHAPTER 15

PERSONALITY AND CHARACTER

While by various names personality and/or character has been a concern of teachers for thousands of years, for most of these years there has been little scientific knowledge of their nature, their development, or their measurement. In the last several decades, however, techniques of psychological and sociological research have been applied to the phenomena and currently there is a considerable body of demonstrable knowledge about them. Enough children now have been studied in enough different ways to permit the charting of trends and patterns in personality and/or character development. And, necessarily, the measurement of personality and character has kept pace with the scientific study of their nature and growth.

In the schools, increased knowledge has been followed by increased specific attention to the development of personality and character. Moreover, curriculums, methods, and reading material are being designed in conformance with the realities of developing child personality.

In regard to measurement of personality, however, the advances of psychologists and sociologists have had less immediate consequence for teachers. Evaluation of personality and character variables has no doubt become a more frequent activity of teachers, particularly elementary teachers, but the evaluations still are gross and the traditional methods of observation and impressionistic rating still are the ones most frequently used. While Buros' *Fourth Mental Measurements Yearbook* lists some 121 standardized instruments for personality and character measurement, only 17 of these would be of use to teachers.

Apparently, there are a number of reasons for this condition. Most of the tests and other devices for group administration are *not* sufficiently reliable to yield any but the most tentative information about individuals. On the other hand, the individually administrable instruments that may be used with more confidence require special training for use and their administration usually demands far more time than teachers have available. A third deterrent to accurate personality—character measurement by teachers is the helter-skelter state of definition for the phenomena and their dimensions. Finally, schoolmen and the public have yet to decide just how much concern the school *should* have for evaluation of personality and character, with the result that teachers may more easily dispense with it than with the evaluation of subject achievement.

Therefore, like the previous one, this chapter is strictly limited in scope. How-to-do-it treatment is given only to those dimensions that many teachers at all grade levels normally evaluate and only to such procedures for measuring them as may safely be used without special training or knowledge. These are thought to be

1. The measurement of attitudes and interests that have educational implications, by means of self-reports and tests of opinion, and

2. The evaluation of certain character attributes or citizenship through observation and analysis of peer opinion.

Other phases of personality or character measurement will be described merely for purposes of orientation to the general processes and problems of the field. The technical training in clinical and case study techniques needed by school psychologists, psychiatrists, psychometrists, social workers, and guidance officials is beyond the scope of this volume.¹

GENERAL CONSIDERATIONS

Definition of Personality and Character

The dimensions of personality are for the most part covert and inferred, for personality itself is an abstraction. Like intelligence, the term is a construct, the name for whatever pattern of consistency may be observed in a person's behavior and, like intelligence, what comprises it or what we call its dimensions is the distinguishing feature of this consistency. For example, we may say that Tom has a bad temper and Joe has an easy one. What we may mean is, first, that Tom behaves consistently in one way and Joe in another and, second, that Tom's consistency has frequent angry outbursts as one of its components and Joe's does not have such a component.

Character similarly is a symbol for observed consistency in behavior and in many contexts it is used interchangeably with personality. Generally, though, it refers to such behavioral consistency as has an important bearing on a person's moral or ethical reputation while personality usually is a more inclusive term and implies no particular social evaluation. For example, a "personality" would not be called good or bad but only adjusted, neurotic, rigid, flexible, and the like; whereas a "character" would be evaluated as good or bad, strong or weak.

¹ Normally, these specialists undertake advanced study in behavioral measurement, using such texts as Gladys L. and Harold H. Anderson, *Introduction to Projective Techniques* (New York: Prentice-Hall, Inc., 1951); J. E. Bell, *Projective Techniques* (New York: Longmans, Green & Co., 1948); R. B. Catell, *Description and Measurement of Personality* (Yonkers: World Book Co., 1946); L. J. Cronbach, *Essentials of Psychological Testing* (New York: Harper & Bros., 1949); David Rapaport and others, *Diagnostic Psychological Testing* (Chicago: Year Book Publishers, 1946, 2 vols.); and, of course, the manuals of protocol for important projective tests.

Dimensions of Personality and Character

Dimensions of both personality and character commonly the focus of tests, questionnaires, and rating devices are outlined in Table 31 along with the forms and procedures of measurement and the evaluative standards appropriate to them. The listing is somewhat arbitrary since there is little agreed-upon definition for personality and character or consensus as to their dimensions.

DIMENSIONS ARE INFERRED

As we have observed, all the dimensions are matters of inference rather than observation. For example, we can *observe* a child jumping away from a snake, a youth using a road map, and an adult saving money to send his children to college. But we must *infer* the fear that made him jump, the belief that the map is correct that allowed the youth to follow its directions, and the high value of education to the father, which prompted his saving. Moreover, the fears, beliefs, and values that we talk about and try to measure usually are inferred from many incidents, not just one, and hence are those things we call abstractions. Finally, because the children we *have* observed exhibit certain behaviors from which inferences as to fears, beliefs, and values are appropriate, we assume that children we have *not* observed have "fears, beliefs, and values" in some manner and to some degree.

The measurability of such inferred and abstract dimensions is discussed in Chapter 2, pages 26–28. It is stated there that clear definition of inferred dimensions is essential and, because they themselves are not observable, that directly related observable dimensions must be found for them. It is on these two knots of definition and related observables that personality measurement frequently is fouled.

FEARS, BELIEFS, AND VALUES

Since attitudes, interests, and character attributes (citizenship) will be discussed by themselves, we may confine ourselves here to fears, beliefs, values, and certain other variables of personality structure. Fears refer, of course, to the items for which the pupil would have fright reactions and the degree to which he would be frightened. Beliefs seem to denote those statements, ideas, concepts, etc., that the individual feels sure about or will act upon. That to which values refer is far more complicated, both psychologically and semantically, but for purposes of measurement it may be resolved into the individual's basic preferences as to goals and means. A simple instance of this is afforded by an adolescent boy as against his adult male teacher in the case of a classmate of the boy cheating on a test. The boy probably would *prefer* to cover for his friend whereas the teacher would *prefer* to discover and punish the misdeed. We would say that the boy's primary value in the situa-

tion was "loyalty to his friends" whereas the teacher's value was "to maintain discipline."

Plainly then, fears, beliefs, and values have some common subdimensions: the identity and number of items to which feelings attach, the valence or direction of the feelings, and the intensity of the feelings. So, in measuring fears, beliefs, and values, psychologists must ascertain what pupils fear, believe in, and value; whether their feelings are to approach or to avoid and just how the approach or avoidance is expressed; and, finally, just how fearful a pupil is, how strong are his convictions, and how intense are his values.

PERSONALITY STRUCTURE

Personality structure or pattern often is an object of measurement for psychiatrists and clinical psychologists. In Table 31 are shown three different types of variable commonly used in describing personality structure.

Polar Dimensions. The first type consists of the designation of a number of polarities with which each personality necessarily is involved; dominance-submissiveness, placidity-excitability, and masculinity-femininity, for example. It is assumed that personalities are ranged from one end to the other on each of the polarized continuums and that personalities may be characterized by stating where they are found on each continuum.

Personality Types. A second approach to describing a personality is to state its type and how clearly it is that type. The "types" illustrated in Table 31, cycloid, paranoid, compulsive, are derived from diagnostic classifications for neurotic and psychotic states. Others (sanguine, morose, dependent, aggressive) are simply the way people have become accustomed to "typing" one another. The idea in the establishment and use of types is simply to notice that there seem to be groupings of individuals so far as personality is concerned and to give a convenient name to each group. Usually the basis for the grouping and for the name is a single prominent characteristic that seems to pervade and color all activities and aspects of the person. Cycloid individuals are either up or down, have extreme emotional swings; sanguine are always hopeful and seeing the bright side; dependent are always clinging to a superior or appealing to authority; etc.

Personality Traits. Each of the two approaches to personality characterization just described deals with the personality as a whole, either classifying it or placing it in respect to some polarized behavioral abstraction. The third type illustrated attempts, on the other hand, to identify the prominent elements of the personality and to indicate their relative strength and interrelationship. These elements may be the very familiar ones we have indicated as separate dimensions in Table 31 (Attitudes, Interests, Fears, Beliefs, Values) or they may be particular to some psychological viewpoint or personality test. Those cited in the table are used with a given projective test but are consistent with current theories of personality motivation. Need domi-

TABLE 31

Outline of Personality and Character Dimensions, Measuring Procedures, and Evaluative Standards

<i>Dimensions *</i>	<i>Measuring procedures †</i>	<i>Forms of measurement</i>	<i>Usual and/or appropriate evaluative standards</i>
Attitudes ‡ (toward school, war, minority races, etc.)	1. Observation and use of recording or rating forms: anecdotal, check list, graphic rating scale, "man for man" rating scale, etc. 3. Self-report on what has happened and how has pupil felt: autobiography, diary, free discourse, dream report, check lists, questionnaires for agreement or disagreement with statements of the occurrence or frequency of occurrence of given actions or feelings, etc. 4. Guided response tests asking for expressions of opinion on verbal items keyed to the dimensions in question: yes or no, forced choice among options, "social distance" items, selection or rejection of items in a list, direct statement of feelings about items	Description-classification and ranking	Range of attitudes found in given group with central tendency usually desirable and extremes usually undesirable Idealized conceptions of teachers, psychologists, sociologists, etc., as to a "best" to "worst" hierarchy for any attitudinal area Age expectancies according to the findings of child development specialists §
Vocational Interests †	1, 3, 4	Description-classification and ranking	The interests expressed by successful practitioners in any vocation vs. those of unsuccessful and/or non-practitioners

TABLE 31 (Continued)

<i>Dimensions *</i>	<i>Measuring procedures ‡</i>	<i>Forms of measurement</i>	<i>Usual and/or appropriate evaluative standards</i>
Other interests † (books, recreation, music, etc.)	1, 3, 4	Description-classification and ranking	A premium usually placed on amount and variety
Fears †	1, 3, 4 2. Product analysis: drawings, imaginative writings, etc.	Description-classification and ranking	As for attitudes A premium placed on fewness and these few being rational and socially useful
Beliefs †	1, 3, 4	Description-classification	As for attitudes A premium on rationality and conformance to the community, state, and nation
Values †	1, 2, 3, 4 5. Free-response tests (projective techniques): Ink blots, interpret pictures, finish stories, word association, open-end sentences, figure selection and arrangement, drawings, finger painting, etc.	Description-classification	As for attitudes Premium on their being Judaic-Christian, democratic and scientifically based, and somewhat flexible As for attitudes

TABLE 31 (Continued)

<i>Dimensions *</i>	<i>Measuring procedures †</i>	<i>Forms of measurement</i>	<i>Usual and/or appropriate evaluative standards</i>
Personality structure (Abstractions used by psychologists and psychiatrists in describing "personality")	1, 2, 3, 4, 5 Guided response tests have little use except for rough screening and for group vs. group differentiation.	Description-classification	Lay and professional opinion as to happy and productive "personalities" as expressed in many popular and professional books; optimum includes what is implied by such terms as mature, integrated, well adjusted, outgoing, etc.; worst, by such words as maladjusted, neurotic, rigid, infantile, cold, etc.
Dominance-submissiveness Placidity-excitability Masculinity-femininity Etc.			Legal definitions of insanity Clinical definitions of psychotic and neurotic states
or Cycloid type Paranoid type Compulsive type Etc.			Developmental norms or expectations for growing children and youth as stated by child development specialists §
or Need dominance Need abasement Need affection Etc.			Other things being equal a premium is placed on being like the predominant personality stereotype for a given group.

TABLE 31 (Continued)

<i>Dimensions *</i>	<i>Measuring procedures †</i>	<i>Forms of measurement</i>	<i>Usual and/or appropriate evaluative standards</i>
Character attributes (synonymous with "citizenship" as used in the schools; abstractions used by educators, business men and people generally to describe aspects of behavior which they admire or decry): honesty, neatness, friendliness, obedience, co-operation, tolerance, self-respect, critical thinking, sportsmanship, leadership, sympathy, etc.	1, 2, 3, 4, 5 6. Peer opinion as to popularity, leadership, reputation, etc., as expressed in sociograms, straight rating, "guess who" techniques, etc.	Description-classification and ranking re specific dimensions	A premium placed on maximum possession of each attribute as ideally conceived Age expectancies as to each attribute considered developmental in character §

* All dimensions must have careful operational definition before measurement is attempted. The items are not mutually exclusive, i.e., attitudes are personality variables, and there is no agreed-upon definition for most of them.

† Attitudes, interests, fears, beliefs, and values have these subdimensions in common: feelings toward objects or ideas, their valence and intensity, and the organization among related feeling-object pairs.

‡ All of any procedure is not necessarily applicable to every dimension for which cited. Procedures vary as to reliability and validity but generally are less valid and reliable than procedures of measuring achievement or intelligence. Reliability and validity must be estimated for a procedure at time of use and used in the light of that estimate.

§ Age expectancies and norms for these dimensions are contained in such books as Havighurst, *Human Development and Education*; Gesell and Ilg, *Child Development*; Olson, *Child Development*; Jenkins, Shacter, and Bauer, *These Are Your Children*.

nance refers to a presumed motivation on the part of an individual to dominate others. His need might be weak or strong and he would be assigned a number indicating its strength. The needs that predominate in a person together with their intensities serve as the structure for a personality according to this view.

Forms and Means of Measurement of Personality-Character

Of all behavioral phenomena personality and character are most restricted to the least precise of measurement forms, description-classification. Extended verbal statements are about the only valid means those who engage professionally in the measurement of these phenomena have to symbolize their status. Classification always is possible but so many are the dimensions of personality and character, that when you have correctly classified the pupil in one way you probably have wrongly classified him in several other ways.

Classification as well as description is, of course, applicable to specific dimensions of personality and character. Children and youth can be assigned numbers or words that stand for a given cluster of interests, beliefs, or values; and arbitrary classifications can be contrived for most of the other dimensions. For example, attitudes toward school might be classified in terms of the age level where they are typical and a child's attitude toward school expressed then as juvenile, adolescent, or adult.

Ranking also is an appropriate form of measurement for those specific personality-character dimensions that are unitary in nature. Children could, with validity, be placed in rank order relative to their interest in athletics, their fear of animals, their honesty, and even their popularity. They cannot and should not be ranked on such mosaic dimensions as attitudes in general, personality structure, or social adjustment.

Scale measures may not be used for personality-character or any of their dimensions, measuring instruments with the essential character of true scales having yet to be devised for them.

The great variety of measuring procedures applicable to personality and character are cited in Table 31. As might be expected, the entire repertoire of behavioral measuring devices is used: observation, product analysis, free response, and guided response, and several techniques are employed that are peculiar to this area only: self-report, peer rating, interviews, projective tests, etc. In ensuing paragraphs the six basic approaches to personality-character measurement shown in the table will be explained briefly. Those usable by teachers in classroom situations will be discussed at greater length in connection with the dimensions to which they are most applicable.

OBSERVATION

The rationale behind observation as a means of personality or character appraisal is simple and obvious. The two, as we have stated, are no more than constructs that serve to explain consistent differences in behavior among

individuals. To observe behavior, then, is inherently the most valid means of measuring either personality or character.

Techniques of observation are applicable to all the dimensions cited in Table 31, but particularly so to character attributes. In fact, appraisals of character by means other than observation are held in suspicion by many educators and psychologists. The evaluations that teachers report for any of the dimensions usually are based on observation. Psychologists and psychiatrists employ special observational devices in analyses of personality structure and diagnoses of personality disorders. Children often are placed in controlled play situations and then observed directly or indirectly, through one-way vision screens. The detailed behavior profiles of infants and small children prepared by Arnold Gesell and now used as norms for child development were based on repeated observations of hundreds of children in a special observation chamber.

Observation may be casual and of "real life" situations or it may be formalized and directed toward contrived situations. In either event, it is necessary to translate what is observed into a permanent record of properly meaningful words or numbers. Recording or rating devices frequently used in school situations are illustrated in Figure 59, and a full discussion of observational rating and recording is given in Chapter IV, pages 52-59.

Though observational procedures have great potential validity, much of this is lost in practice. It takes many samples of behavior to yield accurate inferences about the interests, values, or adjustment they reflect and, too often, only one or two samples are used. Moreover, each observer brings his own hopes and apprehensions into the situation and to a degree observes them or their effects rather than the actual behavior of the child being observed. Finally, if the record or rating yielded by observation is read by another, he may impute widely different meanings to the descriptions and classifications from those intended.

PRODUCT ANALYSIS

A second way of measuring some personality dimensions is to analyze the writings, drawings, sculpture, etc. of a person. Since pupil products differ even when they deal with the same material, it may be assumed that some of these differences derive from personality-character differences in the pupils. It remains then only to identify the "personal" aspects of the products and the personality dimensions they represent.

Product analysis is employed chiefly by clinical psychologists for the diagnosis of personality disorders; paintings and drawings are the principal products used. In addition, the imaginative writings of older children, youths, and adults sometimes are studied for clues to certain psychotic states. While handwriting long has been reputed to "reveal" the personality and while there are reputable persons who have had some success in handwriting analysis, it remains largely the province of the charlatan.

In illustration of drawing-personality relationships consider the case of an elementary grade pupil whose drawings over a two-year period were entirely of trees. Always the trees were bent and blown as if in a storm and always one tree was broken or split. Suddenly, the storm motif was dropped, a sun appeared in the sky in some pictures and, in climax, *a drawing was made in which a split tree had been mended with cement*. This was followed by an occasional drawing of things *other* than trees and finally the boy again had the repertoire of subjects for drawing common for his age.

The boy's history during this period involved the following. His mother had died and his father had grieved excessively but passively. He had not permitted the boy to discuss the dead mother nor to display his own grief and he not only had not assumed the mother's role with the boy but, in his sorrow, had withdrawn even some of his previous fatherly attention. At last, the father remarried and the new mother quickly established a warm relationship with the boy.

Product analysis generally is not a device for teacher use. The interpretation of drawings, writings, and the like requires a more extensive study of both the media and personality structure than is included in the average teacher education program. Moreover, even for specialists the procedure presents many problems. The element thought to have personality significance may actually derive from something else—a limited supply of colors, a too small drawing surface, an affectation of his social group unknown to the examiner, even ineptness in a phase of the production that forces elimination or diminution of some factor. The latter is exemplified in the drawing of hands. Hands are hard to draw and children may omit them for this reason. Yet, according to some diagnostic protocols, the absence of hands or other body members has important significance for a child's sex concepts and tensions. There is need usually for a number of products to exhibit the same characteristic before any conclusion may be drawn, and always it is advisable to seek from another source further support for any conclusion.

SELF-REPORT

Both of the two previous procedures seek to measure personality and character dimensions as they normally are manifested. A third means is to ask the person himself to report these manifestations, to describe his own personality and character or the past events, actions, and feelings that might relate to them. This widely used technique is employed by teachers to gain information about interests, by counselors and school psychologists to assess fears, beliefs, and problems, and by psychoanalysts to diagnose subconscious tensions. The many ways of eliciting self-reports are listed in Table 31 and certain guided response devices appropriate thereto are illustrated in Figure 57.

Self-reporting is as widely criticized as it is used. The procedures are eschewed by experimental psychologists and any devotee of behaviorism

simply on the grounds that they involve introspection. Others condemn self-reporting on two rather obvious and reasonable points.

1. Memory is involved and a person's memory about his emotional states and his feeling-toned experiences is apt to be considerably distorted.

2. Personality, character, and all their dimensions are subject to social stereotypes of acceptance and rejection. Children and particularly youths and adults are aware of these stereotypes and hardly any are willing to report themselves in an unfavorable light.

So self-reporting is apt to produce an inaccurate measure of all personality-character dimensions and to provide an unduly favorable appraisal of any that society evaluates. It is simply that no one is likely to remember just how many bad dreams he has had nor how bad they were and very few of us will admit to acts of dishonesty and cruelty or to feelings of hatred or avarice.

For these reasons, self-reporting, whether by questionnaire, autobiography, or interview, is most useful for dimensions for which society has no strong and distinct acceptance/rejection stereotypes. This leaves the procedure applicable to attitudes and interests; to fears, beliefs, and values, as long as such areas of strong social feeling as sex and religion are avoided; and to non-evaluated aspects of personality structure, placidity-excitability, for example. It prevents much use of self-reporting for measurement of character and of socially evaluated aspects of personality structure, i.e., masculinity-femininity.

GUIDED RESPONSE TESTS OF OPINION AND PREFERENCE

A fourth means of personality-character measurement is even more of an indirect procedure than self-reporting. This is the very familiar guided response test of attitude, interest, personal adjustment, etc. As their rationale, guided response tests of personality or character seem to argue that choices expressed among verbal options in a test situation are indicative of the feelings and concepts that normally guide a person's behavior. The three types of item predominant in such tests are illustrated in Figure 58.

Criticisms very similar to those directed toward self-reports are voiced against the tests. It is far from demonstrable that a pupil's behavior during a test has a necessary and clear-cut relationship to given personality-character variables. Pupils may attempt to choose the response most acceptable socially, even refuting their actual feelings to do so. Having no grade to earn, pupils may "kid" the test or mark it in a random fashion. And even if test performance is determined by personality (as it certainly is to some extent) and is undertaken honestly and conscientiously, there is no certitude that a response means what it is supposed to mean.

A simple example of this point of invalidity is afforded by the first item in Figure 58. "Yes No 1. Teachers are more strict than other people." This item might appear in a test of *attitudes toward school*. If it did, it probably would be keyed to a negative attitude toward school, perhaps to the specific idea that if the pupil says "Yes" he is willing to impute an unpleasant

thing to the teacher and thus has at least one ingredient of a negative attitude toward school. On the other hand, one pupil might say "Yes" because his experience with teachers actually had found them more strict than others. Another pupil might say "No" because of opposite experience. Yet each might have similar attitudes toward school or the first could have the more positive and the other the more negative.

Usually, the designers of personality tests are well aware of these weaknesses in their procedure, and they use various means to offset them. In validation studies, it is demonstrated that test behavior and other behavior are related. Items and scoring keys are incorporated to detect dishonesty and insincerity. Many items are keyed to given dimensions, not just one. Thus through increased sampling the effect of irrational determiners of responses is minimized. But despite these and other precautions of test designers, it is advisable to weigh the validity of a guided response procedure very carefully when it is applied to personality or character dimensions.

Because of suspect validity, guided response instruments generally are used only for group appraisal and for preliminary screening of individuals. For example, before and after instruction in literature it would be appropriate to measure the attitudes of pupils toward certain types of reading, but only to see what mean gain there was and not to mark any single pupil's attitude. Or, a group personality test might be administered to truants to detect which of them warranted further individual testing in regard to possible psychotic tendencies.

PROJECTIVE TECHNIQUES

The free response tests used in personality measurement typically are called projective tests or projective techniques. The reasoning behind their use and the pretext for their name is the assumption that an individual will *project* his own feelings and viewpoints into an ambiguous situation. If the situation is clear and meaningful then the likelihood is that the individual will attempt responses that he thinks will cope with the situation: a true-false test, a questionnaire, a game of cards, being scolded by the teacher are examples. On the other hand, if the stimulations themselves are meaningless, vague, unstructured, as an ink blot, a blank sheet of paper, an indistinct picture, or a ball of clay, and the individual is forced to react to the situation, he must give it meaning and this meaning necessarily will reflect his thoughts and feelings. (See Figure 8 for illustrative pictures, page 71.)

The boy's paintings referred to on page 396 were "projections" of his feelings. An action by an eight-year-old boy observed by one of the authors several years ago provides another vivid illustration of projected feelings. The boy had been a reverse truant, he wouldn't go home from school, and he was being interviewed in an effort to determine the cause. After some conversation about himself and things he liked and disliked, he was given a piece of paper, a pencil, and asked to draw something. He countered with the usual

"What do you want me to draw?" and was told, "Just anything." After a minute of sitting and no attempt at drawing, it was suggested that he draw his family. He drew some stick figures of various sizes in front of a stereotyped house and then stopped. When asked, he identified each of the figures, father, siblings, self, but none was identified as mother. "Where is your mother?" the boy was asked. He went to work again and drew a reclining stick figure below the plane of the others and then scribbled many horizontal lines through the figure. "There's mother," he volunteered, "there's mother, way down in the mud."

In reality, of course, projection is a matter of more or less rather than presence or absence. All behaviors must reflect personality to some extent, if there is any significance in the construct, and no stimulation field is likely to be found with absolutely no inherent significance. Projective measurement involves the use of stimulations that contain as few clues as possible to an expected or appropriate response and the interpretation of responses to see what was projected rather than what was reacted to.

Several projective procedures have been standardized and are used extensively by psychological clinicians. The much publicized *Rorschach* (25), the *Thematic Apperception Test* (23) and the *Make a Picture Story* (27) are among them. These standardized tests and other informal projective techniques usually are employed to diagnose personality disorders. They result in descriptive data only or the most gross classifications and thus have little pertinence for group measurement. Expensive both in time and in the specialized training required to use them properly, the techniques usually are not appropriate for classroom use.

PEER OPINION

Since Moreno published *Who Shall Survive*,² the "disposition" of the group has been studied as avidly as the "disposition" of the individual by many sociologists and psychologists. Out of their study has come a new area of measurement, sociometry. It is not within the scope of this text to deal extensively either with the dimensions of group structure and dynamics or with the techniques for appraising these dimensions.³ However, two of the many sociometric procedures are means to assessing social aspects of the individual's personality or character and they are amenable to use by teachers. Moreover, one of these, the sociogram, can provide quick information about the who-likes-whom structure of a class and such information often is of importance to teachers and guidance officials.

Underlying the use of peer opinion to measure personality or character seems to be the following rationale: a pupil's feelings, ideas, mannerisms, etc.,

² *Nervous and Mental Disease Monograph Series*, No. 58, 1934.

³ For further information on group dynamics and its measurement see issues of the journal, *Sociometry*, and Dorwin Cartwright and Alvin Zanders (eds.), *Group Dynamics, Research and Theory* (Evanston: Row, Peterson and Co., 1953).

are more likely to be expressed to his classmates than to his teacher. Thus peers are in the better position to observe each other's "real" personality. Since children and youth are avidly interested in one another, they do observe one another carefully and critically. They may not be as skilled in unbiased observation and in its translation into words as teachers are, but this lack can be compensated for by proper sociometric techniques. In several studies of leadership and adjustment, peer opinion has been found to be more prognostic of success than any other single factor and always it is positively related to the other indexes.

Peer Ratings. A frequently used but still novel technique for classifying pupils according to gross personality or character stereotypes is the "Guess Who" test pioneered by Hartshorne and May many years ago (16). As illustrated in Figure 52, the gist of the procedure is to give brief descriptions of certain *types* of children or youth and then to ask pupils to name which of their classmates fit each one of the descriptions. In devising the descriptions it is essential to phrase them in pupil language, to make each stereotype distinct from any other, and to have no contradictions within any stereotype. Guess Who tests may be used to classify pupils into rough categories of social adjustment, manners, work habits, honesty, etc.

Here are descriptions of several girls in your class. Can you guess who they are? Write in the name of the girl you think each one is.

_____ She has to have her own way or she will break up the game and is very bossy and selfish.

_____ She is so shy that she seems to be afraid to talk, would never contradict anyone, and gets nervous when anyone pays her some attention.

_____ She is always helping other pupils, never tries to get credit for things herself, talks very well when she is with her friends, but not so well in a large group.

Figure 52. Illustrative Guess Who items used to measure the reputation pupils have with their peers.

Pupils can rate each other directly, of course, just as teachers rate pupils, using appropriate graphic or other kinds of rating scales (see Figure 59). Somewhat finer classifications can be made by direct rating and more dimensions can be covered in a given instrument, but a Guess Who procedure is likely to be the more reliable.

The usefulness of either direct rating or a Guess Who test by pupils is limited by several considerations. Many pupils will be hesitant to rate any of their peers negatively, some will overrate their friends and condemn their enemies, others may feel that it is not their business to evaluate one another, and each will be likely to discuss how he rated the others, hence it is necessary

to have excellent rapport with a class before attempting to use the devices. Confidential treatment of data must be pledged to the pupils and they in turn must be enjoined to keep mum on how they marked one another.

Sociograms. The making of a Sociogram is illustrated in Figure 53. In the procedure pupils are asked to indicate their friendship or admiration preferences (dislike, acquaintanceship, or any other basis for discrimination may be used). From these expressions, tabulations may be made for each pupil and a diagram may be drawn to show the pattern of attraction, admiration, etc., within the group.

Pupils indicate how they like one another by answering questions such as this:

"Put down the name of the boy or girl you would most like to work with on a project."

Their responses may be tabulated on a matrix (chart) like this:

	Tom	Betty	Bill	Helen	Mary	Elsie	Etc.	Total
1. Tom	X		1	1	1			3
2. Betty		X						0
3. Bill			X			1		1
4. Helen				X				0
5. Mary	1				X			1
6. Elsie Etc.		1				X		1

And made graphic on a sociogram like this:

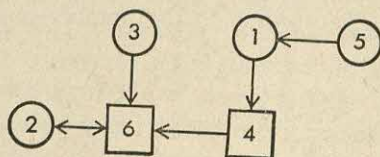


Figure 53. Simplified illustration of the compilation of a sociogram.

From either the tabulation or the sociogram may be derived indexes of popularity, acceptance, etc., for each pupil. The raw tabulations can be converted into rank or classification symbols. In addition, information about the

class as a whole may be obtained from the sociogram. Figure 54 shows the many possible individual and group measures a sociogram may yield.

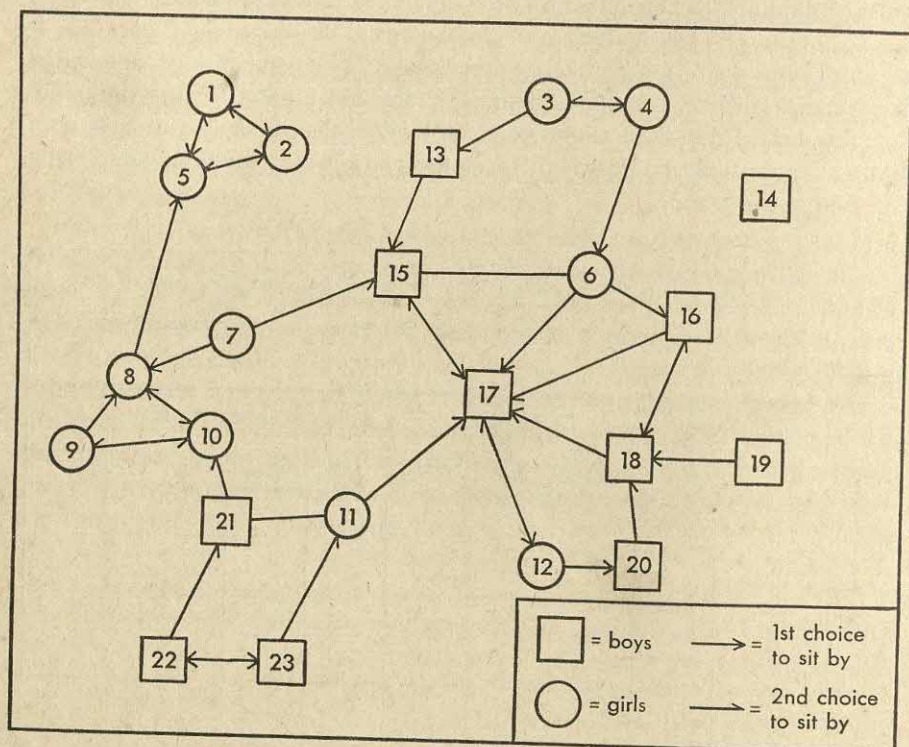


Figure 54. Illustrative sociogram of friendship patterns.

Popularity scores could be determined for each of the pupils, by counting first choices 2, and second choices 1. With such a system, pupil 17 would have a score of 10 and his popularity rank would be 1 whereas pupil 7 and 14 would have scores of 0 and their popularity ranks would be 22.5. Or the pupils might be classified according to the number of first choices they received, as:

4 first	Pupil 17
3 "	Pupil 8
2 "	Pupil 5, 18
1 "	Pupil 1, 2, 4, 6, 11, 12, 15, 19, 20, 22, 23
0 "	Pupil 3, 7, 9, 10, 13, 14, 16, 21

Definite characterizations could be assigned to certain pupils on the basis of the sociogram. Pupil 17 is a leader of both sexes. Pupil 8 is a girl leader and 18 is a boy leader. Pupil 14 is an isolate who apparently has no feelings

of friendship for any of his peers. Pupil 7 is similarly isolated but does indicate a desire for friendship with a boy and a girl.

In addition, the illustrative sociogram might support the following descriptive statements about the class as a whole. The group is organized around pupil 17, the organization cuts across sex lines, and is fairly pervasive. There is one girl clique, pupils 1, 2 and 5, and a male mutual pair. The presence of only one true isolate and the wide distribution of first and second choices indicates a friendly class that probably has been together for a long time.

Individual measures derived from sociograms have the same application as measures derived from other procedures of personality-character appraisal. Obviously, they are particularly useful for such dimensions as popularity, leadership, and friendliness and they can be indicative of social adjustment or maladjustment. Group data have pertinence for group control, seating, committee formation, and guidance activities.

As with peer ratings, the sociogram is a confidential document and pupils will give frank answers only if they trust the teacher. Questions on which the charts are based must always relate to a specific activity or situation *and not to feelings in general*. While the latter may sometimes be inferred from the responses, the use of specifics minimizes the anxiety that pupils may have about their social status in general.

Because the expressions of children and youth about one another are often capricious and are based too much on recent events and because their feelings for one another are subject to change without notice, only the most tentative conclusions may be derived from sociograms. Always other supporting evidence should be required before action is taken. Never should pupils be given "therapeutic" counseling, seating plans be changed, or committees be composed solely on the basis of a sociogram.

PROCEDURES FOR MEASURING PERSONALITY-CHARACTER SUMMARIZED

To summarize this section on procedures, there seem to be six basic ways of measuring the personality or character we impute to individuals because of the characteristic consistency we observe in their behavior. We may look at and listen to what they do (observation). We may study what they have made (product analysis). We may ask them what has happened in their past and how they have felt (self-report). We may determine how they think and feel now (tests of opinion). We may see how they structure an unstructured situation (projective tests). And, finally, we may ascertain their reputation or popularity among their peers (peer opinion). As a rule, projective tests and product analysis require special knowledge and technical skill beyond that possessed by most teachers.

Evaluative Standards for Personality-Character

Few things are more subject to evaluation than a child's personality or character. He makes and loses friends, wins and displeases teachers, dominates

or submits to groups according to the value others place on them. While they usually are not marked in the same sense that subject achievement is marked, many report cards have space for teachers to check their satisfaction or dissatisfaction with the pupils' citizenship (Figure 55). And if they are not

PERSONAL AND SOCIAL GROWTH

Listed below are desirable traits which the school and home seek to develop in children. A check indicates satisfactory growth in the traits listed below.

	1	2	3	4
Contributes to group planning and thinking				
Plans work well				
Completes work begun				
Works co-operatively with others				
Listens well				
Uses time to good advantage				
Is courteous and considerate				
Takes active part in Physical Education				
Shows good sportsmanship				
Practices good health habits				
Observes safety rules				
Respects property of others				

Figure 55. Example of a form to report on citizenship items. (From a report card used in Sacramento County, California, Elementary Schools)

graded per se, evaluations of personality and character often affect the subject marks given by many teachers.

On the other hand, precisely defined evaluative standards are almost totally lacking for the two. Those that are used, as specified in Table 31, tend to be vague, subjective, based on philosophy rather than practicality, and otherwise tend to violate the conditions of valid standards (see pages 193-194). Often evaluations are made without reference to any known standard. A teacher or counselor simply translates his impression of the pupil directly into an evaluative symbol.

Standards in use seem to fall into five principal categories: norms, idealized conceptions, characteristics of criterion groups, laws, and clinical definitions. The first, norms, probably are the basis for most judgments.

NORMS

Norms may be the attitudes found in sixth-grade boys or the average sportsmanship of secondary pupils. Their use involves the assumption that a group mean or one of its extremes epitomizes the status of most value and lesser values accrue to deviations from this point. The most valid instance of norms as evaluative standards may be in the evaluation of social development where usual sequences of development and the range of behaviors usual at any age constitute the norms.

IDEALIZED CONCEPTIONS

The second category of standard, idealized conceptions, is to the educator what pure gold is to the assayer. To the assayer, value increases as the metal contains more gold and less of other elements. To the educator, the value of the pupil's character increases as it contains more of the traits described in the idealized conception and less of other, usually antithetical, traits. An idealized conception of personality or character or any of their dimensions is a description of a condition of maximum value. The description is the "pure"

gold and as pupils seem to approach it they are given high evaluations. An example might be

Co-operation: Always thinks of the group ahead of himself; works as hard on tasks others initiate as on his own; volunteers his talents as needed but withholds them when they are not needed; can change from leading to following according to the demands of the task without ego involvement. Etc.

CRITERION GROUPS

The characteristics of criterion groups constitute a third important type of evaluative standard for these phenomena. The usual process in this type of evaluation is to select a group who epitomize an extremity of value for the phenomenon in question and then to record the characteristics of the criterion group so that any individual may be compared with them. An example of the use of criterion groups may be seen in the evaluation of vocational interests. Successful practitioners are selected in several important occupations or occupational areas and the responses of these persons are obtained to many questions of opinion. An individual is considered to have "an interest" in the occupation in question to the degree that his responses to the same or similar questions of opinion are like the responses of the criterion group.

Standardized interest tests demonstrate a further convention in the use of criterion groups as evaluative standards, namely, the incorporation of a standard into the measuring instrument so that the test score is immediately an evaluative symbol. First, a number of items are prepared that discriminate between the criterion group and people in general. Then a test is prepared and administered to the criterion group. The item by item responses and/or the total scores of criterion individuals are recorded. From these a scoring key or score-interpreting device is prepared that automatically shows how any testee compares with the criterion group. As his item responses or score approximates the item responses or score of the criterion group he is assumed to have an interest in the occupation it represents.

The standards of idealized conceptions and norms often are similarly incorporated into measuring devices so that evaluations may be made or read automatically rather than as a separate step. The rating "scales" used for citizenship or character express evaluations rather than measurements only (see Figure 59) whenever their numbers or intervals stand for range in a norm group or for degrees of approximation of an idealized conception.

LAWS

The use of laws as evaluative standards for personality and character dimensions occurs primarily in connection with determinations of insanity, immoral conduct, and professional incompetence. Both statutes and common law restrict the activity and authority of "insane" persons and many statutes define penalties for certain types of behavior. These laws are the standard of value for courts in judgments about insanity and immorality. In much the

same fashion, states, municipalities, and occupational associations have described the conditions under which licenses may be revoked, tenure covenants broken, and civil service employees dismissed (see Figure 56). The wording of these laws and regulations as to competency, honesty, loyalty, moral lassitude, and the like are the standards that courts and commissions apply to the cases presented to them.

ARTICLE 2. DISMISSAL OF PERMANENT EMPLOYEES

(Education Code, State of California)

13521. No permanent employee shall be dismissed except for one or more of the following causes:

- (a) Immoral or unprofessional conduct.
 - (b) Commission, aiding, or advocating the commission of criminal syndicalism, as prohibited by Chapter 188, Statute of 1919, or in any amendment thereof.
 - (c) Dishonesty.
 - (d) Incompetency.
 - (e) Evident unfitness for service.
 - (f) Physical or mental condition unfitting him to instruct or associate with children.
 - (g) Persistent violation of, or refusal to obey the school laws of the State or reasonable regulations prescribed for the government of the public schools by the State Board of Education or by the governing board of the school district employing him.
 - (h) Conviction of a felony or of any crime involving moral turpitude.
- Etc.

Figure 56. Example of a legal evaluative standard: the dismissal clause in the California teacher tenure law.

CLINICAL DEFINITIONS

Clinical definitions, the final type of evaluative standard for personality-character to be discussed, are the standby of psychologists and psychiatrists. Through experience and research they have established many categories of maladjustment, neurosis, and psychosis. They have found that certain expectancies as to duration and development belong to each of the categories, that given types of treatment are more satisfactory for some than others, and that stereotypes of background and etiology tend to accompany each.

The psychological and psychiatric professions have written in their texts and case books definitions of these many states.⁴ After interviews and tests, the examiner may evaluate his patient's behavioral ailment by naming it schizophrenia, paranoia, hysteria, neurasthenia, or some such. This means

⁴ The definitions are not precise and they shift from year to year and from writer to writer.

simply that the patient's condition has been judged most like that definition bearing the name used. (The definition may be subjective and based on the psychopathologist's experience rather than writings.)

Of the five types of standard described for use in personality-character evaluation, only norms and idealized conceptions are likely to be of use to teachers in their evaluations of pupil behavior. To ascertain and to prepare for use the characteristics of a criterion group is so time-consuming and technical a task as to make the criterion group useful only in standardized testing. Teachers usually do not make legal judgments; and clinical definitions, of course, should not be used by any but graduate psychologists and psychiatrists.

Making Evaluations of Personality or Character

In basing evaluations of personality-character dimensions on norms and idealized conceptions, it is especially necessary to pay strict attention to principles of valid evaluation (see pages 193-205). The norms and the conceptions are expressed verbally and inevitably they involve the bias of the evaluator. There must be safeguards against the errors that derive from either of these conditions and other aspects of the evaluation must be indefectible so that error will not be compounded.

No research is available to indicate the proper way to mark personality-character dimensions nor does the collective experience of teachers offer any consistent precedents. From the general tenets of good marking, however, we may derive a few guides for practice.

1. If citizenship or any personality-character variables are to be graded, mark them separately and as such. *Do not allow them to be a component in any subject or skill mark.*
2. Mark the dimensions of character or citizenship separately. Assign an evaluative symbol or symbols to each of the factors you weigh: neatness, punctuality, honesty, regard for others, etc. *Do not mark character or citizenship as a whole.*
3. Use verbal evaluative statements rather than letter marks where possible.
4. There is no special merit in any particular set of evaluative symbols (*A, B, C, D, F; S, U; S, U, O; E, G, F, P; etc.*), when they are applied to dimensions of personality or character. It is easier, of course, to use a dichotomous or trichotomous set rather than five because there are fewer lines where close decisions are involved. The nature of variation in personality-character dimensions, however, may be better suited to five valued judgments than to two or three.

ATTITUDES AND INTERESTS

A pupil's attitudes and interests are significant to teachers in two ways. First, they affect what and how efficiently he learns and, second, changes in

them often are a specific objective of instruction. It is easily demonstrable that children tend to remember things they like and to forget what they dislike, that boredom makes instruction ineffective, that a pupil's interest in fishing, say, can be exploited in a reading lesson. Moreover, the purpose of schooling is not only to teach facts and skills but also to teach appropriate attitudes and interests.

The terms "attitudes" and "interests" seem to refer to the relatively mild feelings stimulated by the things that flood in and out of a person's experience. In the most prevalent usage the word "attitude" more often is reserved for feelings about things external to the person: school, war, minority groups, political parties, etc. Interest, on the other hand, usually denotes a feeling attached to an activity in which the person may engage: reading, baseball, bridge, painting. An additional difference between the two is that attitudes may range from negative through neutral to a positive valence while interests may range only from a neutral to a positive feeling. Of course, the idea of aversion can be and often is placed as an opposite pole to interest and then the variation is the same as for attitude. Obviously, the two are so allied in meaning that they may be used interchangeably in many situations: "What is your *interest* in school?" and "What is your *attitude* toward school?" mean essentially the same thing. In our treatment we shall use "attitude" where it is the usual term and "interest" where usual, but the means of measurement presented are equally applicable to both.

Dimensions

The more important dimensions of attitudes-interests are

1. The things to which feelings attach,
2. The valence of these feelings, and
3. The intensity of the feelings.

Among the categories of feeling-toned things important to teachers seem to be the following:

School subjects and important parts thereof, such as grammar in English and dates in History.

Teachers

School and its components

Entities or abstractions having cultural significance:

Science

Religion

Country

War

Freedom

Work

Minorities

Etc.

Reading matter
 Recreational activities
 Occupations and their aspects

Measuring Procedures for Attitudes and Interests

Measuring attitudes and interests involves the determination for any pupil of the things that arouse feeling, the intensity of the feeling toward them, and its valence or direction. If the things are known or are not themselves in question, the procedure has to do only with determining valence and intensity.

SELF-REPORTING

In Figure 57 are several illustrations of guided response items used to measure attitudes and interests through self-report procedures. The first sample is possibly the most common type in use. In effect, it asks the pupil to rate the degree and direction of his feelings toward outdoor recreation. The gradations of feeling may be increased or decreased, the manner of statement may be varied, and the item is applicable to anything for which a pupil can or will express his feelings.

	Always	Frequently	Seldom	Never
1. I like to play outdoor games and sports				
Etc.				

Yes

1. I would rather listen to a band than play in one.

No

Etc.

Put yes by the things you like and no by the things you dislike

_____ Swimming	_____ Adventure
_____ Poetry	_____ Romance
_____ Movies	_____ Outdoor things
_____ Comic books	_____ Puzzles
_____ Checkers	_____ Mysteries
_____ Canasta	_____ Animal life
_____ Crossword puzzles	_____ Cowboys
Etc.	Etc.

Figure 57. Examples of guided response items used in getting pupils to report on their own interests and attitudes.

The "Yes, No" item following is again widely applicable but any given item measures direction only, not intensity of feeling. Intensity can be gauged, though, by having several items imply different degrees of feeling about the same thing. Another way is to consider that items indicate equivalent intensity but to assume that the more items are marked in a given way, the more intense are the pupil's feelings in that direction.

The third sample is an inventory or check list of interests and aversions. Feeling-toned objects and the direction of feelings are measured but not intensity.

Total scores may be derived for such guided response self-report instruments. For items of the first type, the degrees of feeling may be weighted 1, 2, 3, 4; -2, -1, +1, +2, or whatever is convenient. A summation may be made of the weights any pupil has indicated by the way he has marked the items. The second type yields a total "Yes" score and a total "No" score or, by counting Yes as 1 and No as zero (or the reverse), a single score can be obtained that means all *No* at one extreme and all *Yes* at the other. If yes's and no's are keyed variously to positive and negative directions, a key must be prepared and the opinionnaire scored in the fashion of a true-false test (see page 98). Total scores for the third procedure may be obtained by counting the checks given to things in the same category. If the items in the list constitute a single homogeneous grouping, the total number of checks is the score.

Guided response self-reporting instruments are amenable to preparation and use by teachers for the measurement of attitudes and interests. In addition to adhering to general rules for instrument construction (see Chapter 6, pages 90-118) it is well to observe the following cautions.

1. Ask only for feelings of which pupils are likely to be conscious.
2. Expect no meaningful results if items relate to areas of strong and stereotyped group feeling, i.e., sex, religion, morality.
3. In extracting total scores, add only the items which refer to the same category of object and to the same valence of feeling (unless differences in valence are accounted for by weighting).
4. Rank numbers or classification symbols are the only derived scores appropriate to self-reports of attitudes and interests.
5. Key sufficient items to the attitude or interest object in question to appraise it fairly. While no general rule is possible as to the minimum number of similarly keyed items needed, certain considerations have broad applicability. Complex feeling-toned objects need more items than simple ones (i.e., interests in *reading* would require more items than attitude toward a *given book*). Guided response procedures usually are less reliable for the measurement of feeling than for the measurement of achievement, hence longer instruments are indicated to compensate for this. If a given object for which

feelings are to be gauged has several distinct parts or phases, one or more items should be keyed to feelings toward each of these.

6. A pupil's responses to each item in themselves constitute descriptive data about his feelings and total scores often may neither be desirable nor particularly meaningful.

Less precise numerically but as valid potentially, measures of interest and attitude may be obtained by free-response self-reports and by interview. These procedures will produce only descriptive or at best classificatory data and they are likely to suffer from incomplete coverage of any attitude or interest area. Pupils' free writing and talking usually is unsystematic and the first things thought of are likely to be exploited at the expense of others equally important in the pupil's life. A series of leading questions for writing and a carefully structured interview can overcome some of this weakness.

TESTS OF OPINION

In self-reporting the approach is direct: "What have been and what are your attitudes?" The second method of use to the teacher is less direct: "What are your opinions about certain things? From these I may infer the nature of your feelings past or present."

In Figure 58 the three most widely used types of guided response item are shown as applied to the measurement of opinion: true-false, multiple-choice,

Yes	No	1. Teachers are more strict than other people.

_____ 1. If I were on the staff of a newspaper I would prefer to:		
		a. Write a humor column
		b. Write editorials on social issues and politics
		c. Be the crime reporter
		d. Write feature stories on interesting people and unusual events

Suppose that you knew people in each of the following occupational categories. What relationships would you like to have with them? Indicate this by matching one or more numbers from the the right-hand column with each occupation in the left-hand column.		
_____	Lawyer	1. Perform a service for you
_____	Bus driver	2. Speak to in a friendly way
_____	Teacher	3. Invite to your house
_____	Sailor	4. Ask to join your club
_____	Store clerk	5. Be your best friend
_____	Loan shop operator	6. Marry
_____	Etc.	

Figure 58. Examples of guided response items used to measure pupil opinion.

and matching. The only difference between their application here and to achievement (see pages 106-109) is that pupil responses purportedly are based on feelings rather than on knowledge or skill. As with their use in achievement, each item must have definite keying. In this instance, the keying is to a given feeling-toned object, to the direction of that feeling and, perhaps, to its intensity as well.

Agree-Disagree Items. The way in which the sample items in Figure 58 are keyed may serve to illustrate the process. The first item is keyed to a pupil's attitude toward teachers. The assumption is made that pupils who dislike teachers are apt to magnify their strictness. Consequently, an item is devised for agreement or disagreement which expresses an exaggerated view of the relative strictness of teachers. A *yes* response to the item means that the pupil agrees with the exaggerated opinion and, hence, according to the assumption, dislikes teachers. A *no* response means that the pupil rejects the extreme view and, hence, probably does not dislike teachers. This and other yes-no items are keyed to object and valence but not to intensity of feeling. If many items are keyed to the same thing, total scores can be considered indicative of intensity just as were total scores for yes-no self-report items.

Paired Comparisons. The second of the items in Figure 58 illustrates *comparison keying*. Here, the pupil is required to compare his feelings for each of the four activities stated as options. His choice presumably indicates that he likes the activity chosen better than any of the others. Each of the activity options represents or is keyed to a different type of interest category; thus, the option that a pupil elects is indicative of his predominant interest category. In the illustrative items these might be a. fun, b. ideas, c. excitement, d. people.

Many standardized interest and value tests employ this type of item. Generally, they force many comparisons of each interest or value in question with every other. Hence, scores may be indicative of intensity of feeling as well as valence and object, intensity being inferred from the number of things over which the thing in question is preferred.

Matching Items. The last item in Figure 58, the matching one, is a particularly flexible and efficient type. Several objects of feeling can be equated with as many or more expressions that indicate directions and/or intensities of feeling. In the one shown, an example of the so-called social distance device first conceived by Bogardus, the testee is asked to state the relationship he would be willing to have with persons in different occupations. Each relationship option varies in intimacy; it is assumed that more intimate relations are offered only to persons toward whom one has strong positive attitudes; therefore, the item is keyed to valence and intensity of attitudes toward occupations. The matching type of attitude-measuring item, particularly the "social distance" variant, frequently is employed by sociologists in surveys of the prestige or status value of different occupations, nationalities, religions, etc.

Selecting Items. The selection of items for opinionnaire measurement of attitudes and interests is more complicated than the selection of self-report items. There it is simply a matter of asking efficient questions about how the pupil feels or has felt. In the present case, however, stereotypes of opinion must first be determined that are thought to represent various feelings and items must then be phrased to reveal which stereotype any respondent prefers.

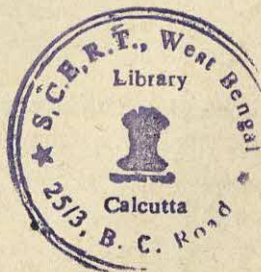
The sources of the stereotyped opinions are groups of people who may be assumed to have given feelings. For example, it could be assumed that truants dislike school and that honor pupils like it. The typical statements of opinion made by each group about school would then be the stereotypes from which test items might be devised.

Scoring Tests of Opinion. Guided-response tests of opinion are scored much as are guided response tests in general (see page 98). A key is needed, and the papers are marked according to the key. Rather than being correct or incorrect, options are indicative of different valences and intensities of feeling.

If attitude or interest in only one thing is being measured, a single score may be derived by counting only those responses that display a given feeling toward the thing in question. The first example in Figure 58 is a type of item often found in interest-attitude tests with such a single focus of measurement, yielding a single score. Pupil responses to a number of these items would be credited or not according to the key and those credited added to give the pupils' scores. Thus, in a test containing items like the one in the illustration, a pupil might have a score of 27 for his attitude toward school, another might have 35, and a third, 40. This should mean that the third pupil liked school better than the second, the second better than the first.

The second item in Figure 58 is a sort that is often used in tests with a multiple focus of measurement. Through one test there is an attempted appraisal of feelings about many things. Obviously, single scores are obviated for such tests and as many scores must be obtained as there are different feeling-toned categories involved. The illustrated item might be from a test of "predominant interests" and a separate score should then be obtained for each category of interest involved in the test. Each item might or might not have an option keyed to each category, but all the responses that favored "fun" would have to be added, all that represented "ideas," "excitement," "people," and other categories would have to be added and the score for one pupil's paper would be a series of numbers, as for example,

Fun	—	13
Ideas	—	17
Excitement	—	4
People	—	20
Etc.		



Such scores would be interpreted to mean that a pupil's interests were more in the areas having high scores than in those having low scores. The scores used as an example show that the pupil was slightly more interested in people than in fun or ideas and was little interested in excitement.

Significance of Scores. Because the scores yielded by tests of opinion are based on items whose equivalence or relative significance is indeterminant, *they may be considered indicative of rank order only.* Moreover, the rank indexes they yield necessarily have a relatively high standard error of score (see page 171). Hence, conclusions should be based only on wide differences in rank and even then conclusions must be considered tentative and gross, much less dependable than those based on achievement and intelligence tests. The extent of rank differences needed for conclusions about interests and attitudes is shown in the Kuder Preference Record (Science Research Associates) where only scores above the 75th percentile and below the 25th percentile are deemed significant of interest on the one hand and aversion on the other.

Sometimes measures are not desired for each pupil's opinions but rather for group opinion with regard to certain things. While the third sample in Figure 58 is useful for measuring a given pupil's attitudes, it also is appropriate for measuring group attitude toward occupations as well. Group attitude is the determiner of prestige so, in effect, this social distance measurer can gauge the prestige value of various occupations. Scoring papers for this purpose involves tallying the way all the tested persons respond to each item, rather than how each person responded to all the items. We derive *occupational scores* rather than *pupil scores*. This type of scoring can yield information of value about books, games, assignments, etc., as well as occupations.

It is expected that teachers can and will devise their own tests of opinion to gauge pupil attitudes and interests. Many well-designed standardized tests are available, of course, and these may be employed whenever they suit the instructional or guidance needs of a particular school situation. The observations just presented may be helpful in the construction of tests of opinion, as may be those made for self-report instruments since the two have much in common. In addition, it is well to pay strict attention to general rules for the construction of guided response instruments (see pages 90-118). In devising and using tests of opinion it must be assumed that they are less valid and reliable than the tests of achievement that the same teacher has devised. Because of their susceptibility to insincere and false answers, no conclusions should be drawn without corroborative evidence.

Standards in Attitudes and Interests

As a rule, marks are not assigned to pupils' attitudes and interests, even when they relate to a given subject, but only to achievement in that subject. Hence, evaluative standards have little practical significance for teachers. The one major exception to this generalization is in the case of vocational interests.

Vocational interests must be evaluated because they are necessary factors in vocational preparation and the secondary schools have undertaken some degree of vocational preparation for youth. The applicable standards are, of course, the status of interests for successful practitioners of different vocations and the interests necessary to enjoy or at least to tolerate the training necessary for the vocations. The former have been formalized in the norms of published standardized tests and are applied automatically in the testing process. The latter are nearly all subjective standards, ideas in the minds of teachers representing their experiences and surmises.

EVALUATING CHARACTER AND CITIZENSHIP

The other aspect of the personality-character phenomenon for whose measurement teachers have particular concern is the matter of character itself. It used to be the term used to denote the total individuality or characteristic state of being for a given human, but for this broad signification it has largely been replaced by the word personality. As we stated earlier, *character* now is used most often to refer to that which other persons evaluate from a moral and ethical point of view. Its dimensions, therefore, have no necessary psychological import but are simply the terms applied to aspects of behavior that may be judged desirable or undesirable, right or wrong, good or bad.

Dimensions of Character or Citizenship

Some of these attributes were listed in Table 31: co-operation, critical thinking, friendliness, honesty, leadership, neatness, obedience, self-respect, sportsmanship, sympathy, and tolerance. Others could be added almost ad infinitum and no listing may be considered definitive. What the words symbolize may overlap and they represent neither a logical nor a psychological system. But they *are* qualities that parents and teachers want children to acquire or to exhibit, and what they stand for affects the welfare of the class and the school.

Citizenship as used in most school situations is a counterpart, if not a synonym, of character. Actions with respect to other pupils and adults and with respect to the legal and ethical norms of society seem to be the referents of citizenship so there is little difference in operational definition between the two. Character may relate more to what underlies ethically judged behaviors and citizenship, somewhat more to the behaviors themselves, but for our purposes we shall treat them as one.

Dimensions of character-citizenship must be selected to suit a particular situation. Such factors as honesty, co-operation, and obedience are likely to obtain, of course, for nearly all situations, but others less basic may not, i.e., tact, altruism, etc. In planning the dimensions of character or citizenship you wish to evaluate it is well to consider the following.

1. The dimensions are abstractions and hence matters of inference. For example, "honesty" does not stand for any given action, thing, or process but rather for the fact that certain actions seem to have something in common. We name the existence of a given common attribute "honesty"; whenever we attempt to measure it we may only infer its "existence" from the behaviors we observe.

2. Most of the dimensions likely to be appraised in the name of character or citizenship are considered to be attributes of behavior inferred to exist in some degree from zero to maximum. Thus, for the majority of them the most important subdimension is amount or frequency: how obedient or how often obedient, when co-operative or how much co-operation, etc.

3. Because they are to be inferred, they should be defined in terms of the actions or action situations from which given inferences are to be drawn. An illustration of such definition is this one for obedience: responds quickly to orders, follows directions exactly, anticipates what the teacher wants, and acts accordingly.

4. Because character and citizenship dimensions are qualities of learned behaviors, age differences are apparent for them. Certain dimensions may be valid at given ages but not at others; only those dimensions should be appraised at a given age whose possession is possible for that age pupil. For example, tolerance is an inappropriate dimension for preschool pupils but obedience is an appropriate one. Co-operation is hardly a valid consideration until the third grade or so and such attributes of maturity as altruism, idealism, and tact may not be significant until the secondary grades.

Forms and Means of Measurement

Having only inferred and ill-defined dimensions, the character or citizenship of a pupil may be described, roughly classified, or assigned a rank within a group. It cannot be expressed as a scale number and probably the most valid measure is simply a description. In this description, status within certain dimensions may very well be appraised in terms of a class or rank designation.

Observation and peer rating are the only procedures currently having much validity for assessing character or citizenship. Measures based on self-reports and tests of opinion and knowledge have been found to bear a nearly negligible relationship to pupil behavior (21:126-134). Product analysis is hardly pertinent to their measurement and while projective procedures have been found to be of some use in character analysis, their administration usually is beyond the resources of teachers.

BEHAVIOR RATING SCALES

The obtainance and use of peer opinion is described in a previous section of this chapter, pages 399-403, and techniques of observation are treated in detail in Chapter 4. It should suffice here to explain several devices for rating character dimensions. Three such are illustrated in Figure 59.

Graphic Rating Scale					
Honesty	1	2	3	4	5
	Lies and cheats whenever it will help him		Usually is honest but may lie or cheat if threatened or frightened		Always is truthful and honest even when he may be hurt thereby

Modified Man for Man

In appearance and dress most like:

- ____ 1. Tom, whose hands and face are often dirty, hair unkempt, clothes rumpled and soiled, shoes scuffed.
- ____ 2. Bill, who usually has clean hands and face, combs his hair and has pressed clean clothes but who is inattentive to his appearance, is apt to wipe his hands on his trousers and to have his shirt tail out most of the time.
- ____ 3. Jack, who always is freshly scrubbed and combed, whose shoes shine and whose clothes are in perfect repair and press and who takes care not to get dirty or mussed.

Check list

Words checked characterize the pupil in question

____ Courteous	____ Neat	____ Irritable
____ Thoughtful	____ Prompt	____ Selfish
____ Fair	____ Obedient	____ Rude
____ Sympathetic	____ Rowdy	____ Disobedient
____ Friendly	____ Inconsiderate	____ Late with work
____ Cooperative	____ Takes advantage	____ Cheats
		Etc.

Figure 59. Examples of three types of devices used in rating personality-character variables after observation.

Graphic Scales. The form at the top is called a graphic rating scale. It consists of a line with numbers or subdivisions together with brief descriptions of how pupils act who possess different degrees or amounts of the dimension in question. The descriptions correspond with subdivisions, points, or numbers on the line. Pupils are observed and a determination is made of which descriptive statement any one most nearly approximates. A check is placed at a point or number on the line corresponding to the proper description or to a point or number a proper distance from it. The device is useful for all character-citizenship dimensions and, if the "graphic" statements are carefully drawn to cover the variation likely to be observed in the dimension, it is a valid procedure easy to prepare and to use. A single page can contain a number of dimensions and provision for their rating.

Man to Man Scales. The next form, "modified man for man," is really a variant of the graphic rating scale. It permits, of course, only specified ratings and not intermediate ones as does the graphic rating. The form derives

from efforts to measure character or personality by selecting a series of prototype individuals known to those who are to make the evaluations; the evaluators appraise other individuals by selecting the prototype which each most resembles. As usually practiced now and as exemplified, the "man" with whom any pupil is compared is fictional, the names are there only for verisimilitude, and it is just a special type of descriptive rating scale. Its special merit lies in the more detailed description usually given. Its principal flaw is its lack of intermediate ratings but this can be remedied. Obviously, many more than three "men" may and often should be used for the rating of a given dimension.

Check Lists. The last device in Figure 59, the check list, is a crude but widely used means of indicating the attributes most prominent in a given pupil. No continuous variation can be expressed for any of the dimensions and, knowing their inferred nature, the device is inherently invalid for this reason. Variation can be shown by using numbers from, say, 1 to 5 for each factor rather than a check or the omission of a check. Without some index of variation for each of the dimensions, the form serves only to record the observer's gross impressions. It is tantamount to measuring a mountain range by enumerating the peaks that exceed a given height. You would get an impression of an extensive or small, a high or low range but that is about all. Just so, a check list such as is illustrated might distinguish a very pleasant pupil from a very unpleasant one but hardly more.

DESCRIPTIVE RECORDS

Rather than rating pupils' character dimensions, anecdotal or descriptive records may be kept. These have the advantage of greater possible attention to individuality and to variations other than in amount but may lack consistent coverage of all important dimensions for all pupils. Moreover, purely verbal descriptions make for difficulty in comparing pupils. The matter of coverage can be handled by having printed on the record form all the dimensions to be appraised. However, it is thought that a combination of anecdotal record and dimension-by-dimension rating procedure should constitute a more efficient procedure than either used alone.

Evaluative Standards

Idealized conceptions and developmental norms serve as the usual standards for evaluations of character and citizenship. Idealized conceptions, as we have presented them, are statements of things as they should be or of gradations from some lesser condition to this point. When applied to citizenship or character, they are stereotyped descriptions of the maximum manifestation of any trait or this plus stereotypes for lesser conditions (notice the first scale in Figure 59). Developmental norms, on the other hand, would be the character-citizenship attributes typical of children of different ages. If the two are used in conjunction, we have idealized conceptions at each age level based

on the possibilities for children at that age. Such a joint standard is more useful than either of them alone.

The ideal conceptions and the age modes are stated in some elementary grade courses of study and in curriculum manuals, particularly those that advocate "unit" instruction.⁵ Child development textbooks contain developmental norms for many dimensions. More often, though, their determination is left to the individual teacher and, as a consequence, standards for character and citizenship tend to be highly subjective and to vary capriciously from teacher to teacher.

We noted elsewhere in the text (page 191) that measurement and evaluation often are coalesced in practice and that unwitting error may accompany the process. No aspects of educational evaluation are more susceptible to this condition than these we are discussing. In Figure 59, the statements in the first illustrative rating device imply value as well as describe possible status. A child who is rated 5 in honesty not only has been "measured," he also has been judged to be of high worth, for in our society a high degree of honesty is considered a virtue. Too, in the check list at the bottom of Figure 59, certain terms have laudatory connotations and others derogatory. Consequently, the pupil for whom the good words are checked has been praised as well as measured and the one receiving checks by offensive words has, in effect, been blamed. Some mingling of the two operations is close to inevitable because the measurement symbols are nearly always words and the words applicable to character and citizenship frequently have evaluative significance.

Consequently, certain precautions should be observed. The registration of a rating always should follow observations and, if these are many and occur over an extended period of time, ratings should be based on observational notes. The rater should be aware that he is evaluating and not just measuring. The validity of the rating scale as a proper evaluative standard should be established and not just assumed (see Chapter 9, pages 193-194).

Beset with subjectivity and with the coalescence of evaluation and measurement, evaluation of character and citizenship encounters another difficulty in middle-class—lower-class value differences. As we know, the conceptions of character and citizenship that constitute teachers' standards tend to be middle-class in character. On the other hand, a large percentage of pupils in many classes are from lower-class homes and their character may be judged by very different standards. Consequently, many pupils are foredoomed to fail in citizenship simply because parental approval and reproof have been given to the "wrong" attributes. For example, too frequent fighting usually is judged by middle-class teachers to be a sign of poor citizenship. Yet, in many lower-class groups, ability and willingness to fight are highly prized.

⁵ For a detailed examination of citizenship evaluation as it is practiced in the schools of one state, see "Evaluating Pupil Progress," *California State Department of Education Bulletin*, XXI, No. 6, April, 1952.

Summary

The dimensions of personality and character are such things as attitudes, interests, fears, beliefs, and values; the variables of personality structure according to a given theory (dominance-submissiveness, masculinity-femininity, for example); and various attributes of character (honesty, neatness, obedience, etc.). Teachers normally are directly concerned only with the measurement of attitudes and interests with educational implications and the evaluation of certain character attributes, or citizenship.

For attitudes, the more appropriate measuring procedures are observation and the use of rating scales; self-report, either through discourse, autobiography, or some guided response device; and guided response tests of opinion. Judgments about the value of given attitudes usually are based on group norms, on the idealized conceptions of sociologists, psychologists, and teachers, and on age expectancies as described by child development specialists. The measurement and evaluation of interests proceed in much the same way, with the addition of the interests of successful practitioners as the standard for evaluating vocational interests.

Fears, beliefs, and values are approachable through the procedures of observation, self-report, and opinion expression and, to some extent, through product analysis and projective testing. The latter procedures are important for clinical examination of personality structure. In such examination interviews and case studies also are basic procedures. Group tests have little use in the analysis of personality structure save for purposes of preliminary screening. The evaluation of fears, beliefs, values, and other variables of personality structure is based on a variety of standards, principal among them being Judaeo-Christian ethics; community, state, and national customs; and books by professional and nonprofessional persons about "the happy and productive personality."

Character-citizenship dimensions have a psychological significance different from the dimensions cited above. They have to do with reputation as much as with being. So, in addition to the procedures described, they may be evaluated through analysis of the feelings of peers about one another. Sociometry is the term generally applied to this type of measurement and sociograms, "guess who" techniques, and straight peer ratings are among its devices.

EXERCISES

1. Differentiate among the following basic procedures for measuring personality dimensions as to validity, ease of use, and pertinence to particular dimensions:

- Observation of behavior
- Analysis of products

Self-reporting questionnaires
Tests of opinion
Projective tests
Sociometry

2. Inspect several published group personality tests and write a critique of each in terms of its susceptibility to insincere and deceptive answers.
3. Prepare a sociogram for a group of children or youth.
4. Prepare a plan for evaluating the citizenship of pupils at the grade level in which you specialize.
5. Construct a descriptive or graphic scale for rating honesty, co-operativeness, and aggressiveness.
6. Devise a form that a guidance official might use for case studies of "problem" pupils.

BIBLIOGRAPHY

1. Ausubel, D. P., Schiff, H. M., and Zeleny, Marjorie P., "Validity of Teachers' Ratings of Adolescents' Adjustment and Aspirations," *Journal of Educational Psychology*, 45:394-406, November, 1954.
2. Bauernfeind, R. H., "Measuring Children's Strength of Response to Attitude Items," *Educational and Psychological Measurement*, 15:63-70, 1955.
3. Berdie, R. F., and Darley, J. G., "Student Personnel Work—IV. Diagnostic Techniques," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 1305-1312.
4. Brieland, D., "Variation of the Guess Who Technique for the Study of the Adjustments of Children," *Journal of Educational Research*, 45:385-390, January, 1952.
5. Butler, J. M., "Use of a Psychological Model in Personality Testing," *Educational and Psychological Measurement*, 14:77-89, 1954.
6. *California State Department of Education Bulletin XXI*, No. 6, April, 1952. "Evaluating Pupil Progress."
7. Cattell, R. B., and Gruen, Walter, "The Personality Structure of 11 Year Old Children in Terms of Behavior Rating Data," *Journal of Clinical Psychology*, 9:256-266, July, 1953.
8. Clifton, Donald *et al.*, "A Projective Technique for Measuring Positive and Negative Attitudes Toward People in a Real Life Situation," *Journal of Educational Psychology*, 43:273-283, 1952.
9. Eaton, J. W., "Experiments in Testing for Leadership," *American Journal of Sociology*, 52:523-535, 1947.
10. Frank, L. K., "Projective Methods for the Study of Personality," *Journal of Psychology*, 8:389-413, 1939.
11. Furst, E. J., and Fricke, B. G., "Development and Applications of Structured Tests of Personality," *Review of Educational Research*, 26:26-55, February, 1956.
12. Gesell, Arnold, and Ilg, Frances, *Child Development*. New York: Harper Bros., 1949.

13. Gordon, L. V., "Fakability of a Forced-Choice Personality Test Under Realistic High School Employment Conditions," *Journal of Applied Psychology*, 40:258-262, August, 1956.
14. Green, R. F., "Does a Selection Situation Induce Testees to Bias Their Answers on Interest and Temperament Tests," *Educational and Psychological Measurement*, 11:503-515, 1951.
15. Gronlund, N. E., "The Relative Stability of Classroom Social Status With Unweighted and Weighted Sociometric Choices," *Journal of Educational Psychology*, 46:345-354, 1955.
16. Hartshorne, Hugh, May, M. A., and Shuttlesworth, F. K., *Studies in the Organization of Character*. New York: The Macmillan Company, 1930.
17. Heron, Alastair, "The Effects of Real-Life Motivation on Questionnaire Response," *Journal of Applied Psychology*, 40:65-68, April, 1956.
18. Johnson, Hildegard, "Facts and Fantasy in Vocational Interest and Measurement," *Journal of Educational Psychology*, 47:40-46, January, 1956.
19. Justman, J., and Wrightstone, J. W., "Comparison of Three Methods of Measuring Pupil Status in the Classroom," *Educational and Psychological Measurement*, 11:362-367, 1951.
20. Keislar, Evan R., "An Improved Formula for Scoring Certain Guess-Who Ratings at the Adolescent Level," *Journal of Educational Psychology*, 45:151-160, March, 1954.
21. Murray, Henry A., *Thematic Apperception Test*. Cambridge: Harvard Press, 1944.
22. Olson, Willard C., "Character Education," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 126-134.
23. ———, "Personality," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 806-817.
24. Rapaport, David, Gill, Merton, and Schafer, Roy, *Diagnostic Psychological Testing*, 2 vols., 1945 and 1946. Chicago: Year Book Publishers, Inc.
25. Rorschach, Herman, *Psychodiagnostic Plates, Fourth Edition*, 1921-1945, and *Psychodiagnostics, Fourth Edition*, 1949. New York: Grune and Stratton, Inc.
26. Sargent, Helen, "Projective Methods: Their Origins, Theory, and Application in Personality Research," *Psychological Bulletin*, 42:757-793, 1945.
27. Shneidman, Edwin S., *Make a Picture Story*. New York: Psychological Corporation, 1947.
28. Sims, V. M., "Technique for Measuring Social Class Identification," *Educational and Psychological Measurement*, 11:541-548, 1951.
29. Smith, E. R., and Tyler, R. W., *Appraising and Recording Student Progress*. New York: Harper Bros., 1942. (Vol. III of *Adventure in American Education*.)
30. Soderquist, H. O., "Validity of the Measurement of Social Traits of High School Pupils by the Method of Rating by Associates," *Journal of Educational Research*, 31:29-44, 1937.
31. Stagner, Ross, "Attitudes," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 77-84.

32. Thompson, G. G., and Powell, M., "Investigation of the Rating Scale Approach to the Measurement of Social Status," *Educational and Psychological Measurement*, 11:440-455, 1951.
33. Williams, S. B., and Leavitt, H. J., "Group Opinion as a Predictor of Military Leadership," *Journal of Consulting Psychology*, 11:283-291, 1947.
34. Windle, Charles, "Test-Retest Effect on Personality Questionnaires," *Educational and Psychological Measurement*, 14:617-633, 1954.
35. Zeleny, L. D., "Leadership," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 662-668.

CHAPTER 16

SCHOOL-WIDE TESTING PROGRAMS

So far in this section we have dealt with the evaluation of pupil achievement in given subjects or with the measurement of pupil intelligence, interests, citizenship, etc. In the presentation, attention has been focused on the individual teacher and the pupils he may teach. Now in this concluding chapter of the section and the book, we wish to shift our focus to school-wide uses of measurement and evaluation.

It currently is common in American schools and school systems to have an integrated and centrally administered program of testing. The reasons for this vogue seem to be twofold. In the first place there happens at last to be the necessary wherewithal for such programs. The scientific movement in education has provided schoolmen with some essential tools and insights: for example, "normal" probability tables, statistical procedures and symbols, mechanical and electronic scoring devices, and accurate concepts of individual differences, the relationship between achievement and mental maturity, and the significance of pupil interest and adjustment. Standardized tests are being published in great numbers, in wide variety, and in constantly improving quality. Moreover, the education of teachers is such now as to produce more and more teachers with skill and interest in testing. And, perhaps the most important factor of all, graduate schools are training psychometric, guidance, and research specialists who are capable of planning and administering testing programs. In the second place, most elementary and secondary schools are trying to do things that are facilitated if not made possible by school-wide testing. Among these are personal and vocational guidance, differentiated grouping within classes or among classes, determination of the mean achievement of pupils in given grades and subjects, and controlled experimentation with methods and materials.

In our study of testing programs we shall deal first with the phenomena usually tested. Then we shall discuss the several procedures and processes the program may entail. After this, attention will be given to the various applications or uses to which the results are to be put. Finally, some general tenets will be given for efficient testing programs.

Focal Points of School-Wide Testing

The standardized tests given to pupils in school-wide programs usually are directed toward measuring intelligence, reading ability, achievement in

basic subjects, vocational interests, special aptitudes, and extremities of maladjustment. A few systems may undertake more than these and many are concerned with only a few of them. Intelligence probably is tested with the greatest frequency, with reading in the second position. Special aptitudes and detection of maladjustment are the factors most likely to be exempted from a program.

Intelligence, reading, basic subjects, and vocational interests we have discussed earlier. The special aptitudes of primary concern to schools are those for mechanical, clerical, artistic, and especially musical pursuits. Evaluation of the latter is becoming a commonplace prelude to instrumental musical instruction in elementary and junior high schools.

Systematic attempts to detect present or incipient extreme maladjustment are accompaniments to the schools' concern over mental hygiene. During the last fifty years or so, in fact since there have been scientific explanations for disturbed and asocial behavior, teachers and administrators have faced two unpleasant realities. Many "bad" pupils, truants, vandals, thieves, and even the "shy" and the socially erratic but academically brilliant may have personality disorders and not just ill will toward teachers. Some cases of educational difficulty (reading disabilities, speech faults, test fright, etc.) may originate in or be aggravated by a neurotic condition as well as by low intelligence or low motivation. As schoolmen have recognized these relationships, they have seen the desirability of screening pupils for maladjustment just as they have grown accustomed to checking for reading readiness. Because few reliable group tests of maladjustment have been devised, not many elementary and secondary schools include the item in their testing programs.

Instruments and Procedures in School-Wide Testing

The instruments basic to a school or system testing program are published standardized tests. These may be tests of specific phenomena (intelligence, arithmetic achievement, etc.) or they may be batteries. The usual components of batteries are the fundamental elementary school subjects of reading, arithmetic, language, social studies and science. In secondary schools there is less occasion to use test batteries but several are available, covering, as a rule, the traditional high school subjects: grammar, literature, history, algebra, physics, and chemistry (see Chapter 6, pages 121-126, for the general characteristics of standardized tests and Appendix B for test titles).

Of necessity, most testing must be done in groups and tests designed for group use are the mainstay of any program. However, the more efficient programs provide for some individual testing as well. Certainly for intelligence, reading, and deviate behavior the results of mass testing often may need to be confirmed or amplified. As a rule, individual tests are not used automatically but only on referral. For intelligence, the Binet and more recently the Wechsler Intelligence Scale for Children are standard individual tests. Behavior disorders may be diagnosed by the Rorschach, the TAT, or a few other projec-

tive instruments. Several good diagnostic reading tests are on the market, among them the *Durrell Analysis of Reading Difficulty* and the *Gates Reading Diagnostic Tests*.

Obtaining and administering standardized tests for even one school of any size is a complicated undertaking. For a multischool district, a populous county, or a metropolitan school system it usually is a full-time, continuous activity for one or more persons. In procuring and using the tests certain factors need to be considered: selection, ordering, manner of administration, manner of scoring, handling results, and cost. While variation among schools as to purposes and personnel precludes any exact prescription as to each of these, efficiency may be increased by adherence to a few general principles.

Selection of Tests. Publishers' catalogues furnish the most current listings of tests together with information as to cost, ordering, norms, and administration time. They do not, as a rule, say much about the test's validity, reliability, or standardization population. The latter information is critical, of course, in assessing the merit of any test. The most complete single source of validity data, etc., is Buros' *Mental Measurements Yearbook*. The periodical, *Educational and Psychological Measurement*, may contain reviews of given tests, and journals in specific fields (i.e., *The English Journal*) often report critically on new tests in their fields. If specimen sets of tests are ordered prior to selection (and they *always* should be), the test manual may indicate what indexes of reliability and validity have been obtained in the test.

Technically adequate standardized tests are coming to be the rule whereas once they were the exception. Consequently, any school should insist that any test it orders have certain essential features.

1. A *manual* giving complete information as to standardization population, dimensions which the test may measure, norms, administration, scoring, reliability, and validity. For achievement, intelligence, and reading tests, it seldom is necessary to accept reliability coefficients of less than .90. Discussions of validity should be frank (any test is suspect that fails to discuss its validity, or that makes unsupported assertions about its validity) and should involve some logical analysis as well as statistics about correlations and item discrimination.

2. *Record forms* that allow for easy presentation of part scores. A graphic display called a *profile sheet* is the most common form.

3. *Simple scoring keys* for both manual and machine scoring, if the latter is possible.

4. Provision for *machine scoring* if the tests are entirely guided response select-an-answer types.

5. *Separate answer sheets* so that test booklets may be reused. If machine scoring is possible, both machine answer sheets and manual answer sheets should be provided.

6. *Time of administration* at least ten minutes less than the length of the period in which the test is to be given. As long as pupils are not fatigued, longer tests tend to be more reliable than comparable short ones.

Ordering. Each publisher states his own preference and rules for ordering but the following obtains for most publishers. Five and sometimes 10 per cent discounts are allowed on quantity orders. Ordering well in advance of need and stipulation of date of use will insure that the tests arrive on time. If there is doubt as to just what test materials are needed, describe to publisher what use is intended for the test and he usually will be able to send what is necessary.

Manner of Administration. The *sine qua non* of test administration is to secure maximum rapport on the part of pupils. Usually this is easier to obtain by having the regular teacher administer the test than by using a testing specialist. In addition, though, directions must be followed exactly, emergencies must be handled, and timing should be punctual, so it may be necessary to have brief training sessions for teachers who are to administer the tests and even to exclude certain teachers who seem to find test administration especially difficult.

Use of regular classrooms (if they are amenable to a good testing situation) is considered preferable to use of auditoriums or gymnasiums. While mass arrangements require fewer test proctors, they tend to increase the possibility of pupil misdirection and inattention.

Manner of Scoring. Scoring tests is a clerical task and a tedious clerical task at that. The advantage usually cited for teachers scoring tests is that they will note the errors of pupils and make diagnostic use of the test. Equal opportunity to diagnose a pupil's difficulty is provided by returning scored answer sheets to teachers. In fact, rapid use of a key nearly *precludes* any notice of who missed what. Moreover, insistence that teachers score their own standardized tests has led in many instances to negative teacher attitudes toward the tests and even to diminished usage.

Hence, in the authors' view, tests should be scored first by machines or by self-scoring stencils. If this is not possible, they should be scored manually by clerks. Only if neither of these is possible should it be necessary for teachers themselves to score the tests.

Handling Results. Results of standardized testing should be transmitted as rapidly as possible to the persons who will make use of them. The longer the delay between test administration and knowledge of results, the less significance will attach to them. Even if the testing is for some purpose that does not involve the teacher, the teacher to whose pupils a test was administered will, as a rule, wish to know their scores and would seem to have some sort of "moral right" to the information.

It is advisable to consider that scores on standardized tests are *confidential* data: emphatically so for intelligence and personality tests. Only professional persons should have access to them. Obviously, student clerks should not handle them and parents likely will find a verbal interpretation more meaningful than a standard score or a percentile.

As to what the pupil himself should know of his standing on any standardized test, it is thought that his welfare should be the criterion. If knowing a

score or having it interpreted will help him, he should be informed. If it will not, if it will frighten him, confuse him, make him feel loss of face, etc., he should not be informed. On this basis, pupils probably should *not* be told their Mental Age or IQ. They certainly *should* be told the rate at which they can read and be given some meaningful index of their reading comprehension, their spelling ability, their arithmetic skill, and of any item involved in the teacher's regular day-by-day evaluations. Probably, pupils should not be given the results in terms of grade placement or percentile rank in a general population.

Cost of Standardized Testing. The cost of tests and their administration obviously will vary according to the number and kind used and the extent to which special staff or released time is provided for the testing program. But without question the cost of a standardized testing program is trivial in comparison to the total cost of instruction. For example, to give a general achievement test (reading, arithmetic, vocabulary, grammar, and spelling) and an intelligence test to one pupil four times during a twelve-year period and to administer a vocational interest test once would cost in materials only \$1.65.¹ If an individual intelligence test is administered, this might cost \$15.00, but most pupils would not be referred for an individual intelligence examination. These costs are to be compared with the over-all instructional cost of educating a child for twelve years, which is about \$4,000 in California. In many states the figure is less but in some it is more.

Hence, we feel that *cost* should *not* be a *critical factor* in determining what shall be tested and what tests shall be used. Revised forms of tests should be ordered if current tests are dated or faulty even though old booklets are still usable. A less reliable test should not be chosen over a more reliable one simply because it happens to be cheaper. If on educational grounds it is desirable to include six things in the testing program rather than four and to have biennial testing rather than triennial, there should be little hesitation just on the ground of increased cost. In our example, use of two tests at three-year intervals and one once during the twelve-year period required only \$1.65 worth of materials per pupil. If this were increased *tenfold*, the cost would be only \$16.50 for twelve years, or but \$1.38 a year. And this presumes a new test booklet for each pupil for each administration, which of course would not be necessary.²

Locally Devised Tests. A school or system with one or more trained psychometrists on its staff may wish to devise its own tests. The chief advantage of this is that the tests may then be tailored to fit exactly the curriculum, the pupils, and the purposes of the school's testing program, whereas commercially published tests, being designed for general use, are only ap-

¹ California Test Bureau, 1956 catalogue. Other tests would cost a little more or less.

² Sacramento City Unified School District, California, estimates the cost of its standardized testing program at 35 to 40 cents per year per pupil.

proximately suited to the program in any given school. The disadvantages of locally designed tests are several. As a rule they will not yield scores that allow comparisons with a general population of pupils. Because they do not have to survive in the market place, they are likely to have more technical imperfections than published tests. To prepare them is a difficult and time-consuming task. In view of this, the psychometrist and the teachers who work with him may grow impatient, short-cut some of the needed analysis, and produce tests less reliable than their published counterparts.

If tests are to be devised by the school or system, they should receive the same careful validation and standardization as the best published tests. Some information about the process is presented in this text (pages 90-126) and more detailed treatment is found in such books as Bean, *Construction of Educational and Personnel Tests*, McGraw-Hill Book Co., 1953; Travers, *Educational Measurement*, The Macmillan Co., 1955; and Lindquist (ed.), *Educational Measurement*, ACE, 1951.

Cumulative Records. If the results of standardized testing are to be used for guidance and for instructional facilitation, each pupil's score on each test must be entered on a permanent record that can be accessible to persons who need the information. Any form is adequate as long as it is durable, readable, and provides for all the entries needed. Many schools and systems devise and print their own. Most frequently test results are recorded on the same card or in the same folder as is used for a permanent record of subject grades, attendance, health, and other guidance data. Several cumulative records are published, among them:

American Council on Education Cumulative Record Folders, Grades I-III, IV-VI, VII-XII, XIII-XVI; 6¢ per copy at any level.

California Cumulative Guidance Record for Elementary Schools, Grades I-VIII; basic folder \$11.25 per 250 copies (A. Carlisle & Co.).

Cumulative Personnel Record, Grades VII-XII; in three forms, card, folder, and envelope; 5¢ per copy for any form (National Association of Secondary School Principals).

Uses of Testing Programs

As we have observed, standardized testing programs are conducted in order to further certain functions or purposes of the school. Perhaps their foremost current uses are in connection with pupil guidance, ability grouping (often called homogeneous or differentiated grouping), and general instructional facilitation. We shall discuss each of these uses briefly and offer some advice re their efficient performance.

GUIDANCE

Guidance activities in schools are designed to help pupils make the best possible adjustment to schooling. They may involve full or part-time coun-

selors who help pupils with academic and personal problems, programing, particularly in senior high schools, home room programs, group guidance sessions, vocational exploration, and many other things. The scores that pupils make on standardized tests obviously are important for several of these items. If achievement test and mental test scores are known, a counselor is in a better position to discuss low grades, tardiness, disorder, etc., with a pupil. Programing and vocational guidance can hardly be accomplished effectively without evidence of the pupil's ability and interests; test scores together with teacher reports constitute more reliable evidence than the latter alone.

Cumulative Records. It is well to observe a few general cautions in using standardized test scores for guidance purposes. Scores should be entered in cumulative records accurately and completely. Only derived scores meaningful in themselves should be used for such entries. Raw scores are meaningless without the distribution of raw scores for the class or grade and class percentiles may be misleading if the class was other than an average one. As a rule of thumb, enter only percentiles or standard scores derived from the test's norms, age or grade placement in subjects and mental ages, and IQ's. Each entry *must* bear the *date* when the test was administered and the *name* of the person who administered it.

Vocational Guidance. Vocational interest tests tend to be less reliable than achievement tests. Moreover, the vocational interests of many adolescents are subject to unpredictable variation. Consequently, it is well to have a broad base for any decision about vocational training. Teacher and parent reports, interviews and aptitude tests, as well as an interest test should be involved in vocational counseling. Interest tests are so available, so easy to administer, and apparently yield such precise findings that counselors may be tempted to rely on them alone.

Referral for Individual Testing. If an individual intelligence test is to be administered to a pupil or if a projective test and diagnostic interview are to be used, the matter ordinarily should be discussed in advance with the pupil's parents. Only rarely will parents object but many will be angered or frightened if they hear of it after the event.

Parents' concern over the social stigma they think may attach to their child being given a "special test" is exemplified in this incident. Certain parents in a small town consented to an individual intelligence test for their daughter, which had been ordered by the teacher on the grounds that she thought the child was mentally retarded. The teacher called the mother and told her this, by the way. Over an intervening week the mother and the father were extremely agitated. The grandmother was called long distance and drove at once (150 miles) to be with her daughter in this time of distress. The examiner, one of the authors, was asked not to go to the school as he usually did but directly to the pupil's home. The mother then went to school and took her child home on some other pretext so that no one would know that a "psychologist" had looked at her. Fortunately, it was found that the girl was

bright enough but that a combination of an ill-trained and overly rigid teacher, an aggressive older sibling and parental criticism had produced a state of acute anxiety in the girl, with accompanying postural and cognitive difficulties. A change of school, better handling at home, and the amazing capacity for recovery that children seem to have relieved the situation, but the family probably will remember its trauma for a long time.

ABILITY GROUPING

The need to accommodate school programs to the wide range of general and special abilities possessed by pupils is an accepted fact. No longer is there any public advocacy of the "Procrustean bed" approach to compulsory education³ nor is it often found in practice. The most effective method of adjusting programs to individual differences, however, has yet to be found and each of the many procedures now in use has weaknesses as well as strengths.

One of the more widely employed devices is that of "ability grouping." This approach, often called homogeneous grouping, differential grouping, or sectioning, involves the division of the pupils in a grade into several classes, each of which is more "homogeneous" with respect to one or several dimensions than is the whole grade. Intelligence, reading, and/or special aptitude for a subject are the usual dimensions on which division is based, three groupings is the usual extent of division, and the device seldom is used below the seventh grade.

Ability grouping is a highly controversial issue and it is not within the scope of this text to analyze it, let alone attempt to resolve it.⁴ We shall restrict ourselves to a few general comments and then describe how standardized testing may be used to determine the composition of sections.

Factors to Consider in Grouping. In general, segregating pupils by ability prior to instruction has been found most effective for those of low intelligence and least effective for those of high intelligence. Modification of method, materials, and even objectives has been found essential, for unless methods, etc., are appropriate it seems to make little difference whether there is a heterogeneous or a homogeneous group. Finally, ability grouping on the usual bases produces groups only slightly less heterogeneous than the original one and homogeneity with respect to one dimension does not insure homogeneity with respect to another, perhaps equally important.⁵

³ The mythological Procrustus would fit wayfarers to his bed by stretching them if they were too short or lopping off portions if they were too long.

⁴ Grouping *within* a class for reading instruction is demonstrably effective and is widely practiced in the primary grades.

⁵ Tiegs (26:262-285) presents a detailed analysis of ability grouping. Some other important sources of information on this subject are Northby (17) and Otto (18); National Society for the Study of Education, *The Grouping of Pupils*, 35th Yearbook, Part 1, 1936, and *Adapting the Secondary School Program to the Needs of Youth*, 52nd Yearbook, Part 1, 1953.

Obviously, if pupils who are most alike with respect to a given subject are to be placed together for instruction, all the factors that bear importantly on achievement in this subject need to be considered. A battery rather than a single test approach thus is indicated. For the school subjects for which ability grouping usually is practiced (English, mathematics, social studies, and science) the use of an intelligence test, a reading test, and a test of readiness or aptitude in the particular subject seem to be mandatory. If only *one* test can be used for some reason, it is thought that this test should relate specifically to the subject in question.

Frequency and Recency of Testing. It is known that children are erratic in their school performance, that they change over a period of months and certainly years, and that standardized test scores have rather large elements of unreliability. Hence, two cautions should be observed. Several intelligence test scores, several reading test scores, etc., should be used, not just one in each category; and consideration should be given to previous marks in similar subjects and to the reports of previous teachers as well as to test results. Secondly, one or more of the tests used should have been administered *recently*, in the last six months as a rule of thumb. The more than occasional practice of using tests administered in Grade VII or VIII as the basis for differentiated grouping in Grade X has serious limitations if not dangers.

Group for Specific Subjects Only. Assignment to any classification should be for *one subject only*. Two or three categories of general ability should not be established and pupils then assigned automatically to the commensurate sections of each subject they take for which grouping is practiced. Possible variation is too great between achievement in one subject and that in another to permit this. For large groups, of course, the mean achievement of slow pupils tends to be low in all subjects and, for brighter pupils, high across the board. But for any individual no such exact uniformity may be predicted.

Review and Verification of Assignments. Finally, there should be frequent review of assignment to given groups and opportunity for verification testing at any time if parents or teacher do not agree with the assignment made by a counselor or principal. These two processes will more than justify their expense in terms of parental and pupil acceptance of an ability grouping program and in terms of more efficient grouping.

INSTRUCTIONAL FACILITATION

Both the guidance and ability grouping applications of testing programs are intended to facilitate instruction, if indirectly. In addition certain direct applications are possible. The range and central tendency of scores on intelligence and achievement tests for a given class are informative to any teacher. They suggest what progress he may expect from a class and in some cases at what level he needs to start. With respect to individuals, standardized test

scores foretell the poor readers, the slow thinkers, the able pupils, and those for whom instruction must be enriched if it is to challenge them.

If test scores for a class are given in terms of national norms, knowledge of them enables a teacher to select materials more efficiently. For example, a teacher may be able to order any of three textbooks, each designed for use in the same grade but differing as to reading difficulty. After inspecting a tabulation of reading scores for the new class, he may decide to use the easy one, the most advanced, or the middle one. Or he may wish to procure some of each. But in any event, his decision can be more rational with knowledge of the test scores than without it.

Gauging Class Progress by Standard Test Norms. A further use of standardized tests is to gauge the progress made by a given class in a given period of time. If the subject(s) in question are those for which tests with national norms exist, a teacher may pretest at the beginning of an instructional interval, retest at the end, and see how much gain has been made. Since the norm tables for the test(s) will show how much gain might normally be expected in this period of time, the teacher may in effect have some basis for judging the effectiveness of his instruction as compared with instruction generally. If this sort of analysis is made, *it is imperative* that recognition be given to the following. All learning in a period is *not* due just to the teacher and neither is all failure to learn. A teacher's objectives may differ from those upon which the test norms are predicated and the norms will be inapplicable to some extent. Unless a class has a range and mean of intelligence comparable to that of the population upon which the test was standardized, direct comparisons with age or grade norms are inappropriate.

Restrictions on Use of Standardized Tests. In connection with instructional facilitation, two *non-uses* should be mentioned. Subject marks and/or promotion should not be based on published standardized test scores. The tests either will not relate to all objectives and aspects of a subject or, if they do, there is a tendency for instruction to be directed toward passing the test rather than toward permanent important changes in pupil knowledge and behavior.

The other *don't do it* is using standardized test scores in an administrative evaluation of teacher effectiveness. For the teacher to do this himself is one thing, and in some cases desirable, but for the principal or supervisor to do it is another. The practice is condemned by all texts in school administration and supervision of which the authors are aware.

Some General Tenets of an Efficient Testing Program

Organization. If a school-wide or district testing program is to operate efficiently and is to furnish valid information for each of the several uses we have discussed, there needs to be a permanent and stable administrative organization for the program. Just what this organization should be, who will

do what and in what order, remains for each school or system to decide on the basis of its own situation and what it wishes to accomplish with the tests. Whatever the situation, though, one person should have clear responsibility for over-all direction of the program. There should be provision for teacher and principal participation in decisions that affect them and concerning which they are qualified to judge. The psychometric staff should be large enough to handle all technical phases of the program. Finally, there needs to be provision for periodic evaluation of the program and for revision as indicated.

Role of Director. The director of the program and/or his assistants should be required to serve as a consultant to teachers and administrators, helping them understand and use the test results properly and even assisting them in devising their own measuring devices. It is particularly ill-advised for him to devote his time exclusively to the administration of tests and a tabulation of scores. School psychologists and psychometrists who construe their jobs this narrowly tend to lose contact with the instructional program they are supposed to serve. Moreover, teachers think of the testing program as being something apart from them and consequently may give less than adequate attention to it.

An additional nontesting responsibility of the program director should be in-service education for teachers. The purpose of the activity is to increase the competence of teachers in administering and using the results of standardized tests and, as well, to increase their skill in devising and using their own evaluative procedures. Among the possible means of conducting such in-service education are bulletins, seminars for teachers in a building, grade, or subject, demonstrations of how to administer tests, and analysis of tests submitted by teachers.

We stated earlier that the results of testing should be communicated rapidly to all who can use them and we wish to reiterate this point. Unless the results are to be applied to the problems of schooling, there seems to be little value in administering them. The idea is not so much to give a standardized test but to accomplish some *given thing* by administering the test. When teachers and principals participate in administration of the tests and there is some semblance of an in-service education program, there tends to be less "filing" of results.

Scope and Frequency of Testing. What tests are to be given and how often is strictly a function of the local school. Practices vary widely and there apparently are no grounds for asserting that any given practice is best. Los Angeles schools in 1950 were using an intelligence test every other year from Grade I on, an achievement battery at two-year intervals from Grade IV to Grade IX, and in high school one comprehensive test of educational development (11). In some other large districts, intelligence is not tested until the third grade and at three-year intervals from then on. It is fairly common minimum practice and probably wise to use an intelligence test in the primary grades, achievement batteries in the intermediate and upper grades, and an

intelligence and reading test at the beginning of high school. The prognostic value of vocational interest tests before the eleventh or twelfth grade is moot but a number of districts administer them in Grade VIII, IX, or X. Group personality tests rarely are used below the high school level.

A pupil's achievement in a subject may change rapidly in a short period of time, so there seems to be no point of diminishing returns for frequent achievement testing. Intelligence testing, on the other hand, is directed toward an aspect of the pupil that changes slowly and regularly. Moreover, a pupil's status relative to his peers tends to remain fairly constant (see page 380). So there is for intelligence tests some maximum frequency of administration beyond which it may be fruitless to go. Research is lacking to establish just what this maximum is but we suggest that use of intelligence tests at two-year intervals is about as often as needed. More frequent intelligence testing for a special purpose (verification of a possibly spurious test result, change of schools, etc.) is, of course, necessary and worthwhile.

Proper Emphasis. As a final tenet of efficiency for a testing program, the program should be such as to stimulate better instruction and provide important information. *It should not be employed to check on teachers*, to control the curriculum, or to make unqualified comparisons between Teacher *A* and Teacher *B* or School *A* and School *B*. Positive applications of the program are likely to suffer if these negative ones are made. Moreover, effective supervision, curriculum control, and curriculum evaluation involve far more than standardized test results.

Summary

School-wide testing programs usually are directed toward one or more of the following: intelligence, reading, general achievement in basic subjects, vocational interests, special aptitudes, and detecting extremities of maladjustment. The basic measuring instruments employed are published standardized tests that may be administered to many pupils at one time. The use of individual tests is restricted in most schools to pupils who show an extreme deviation either in intelligence or personality.

In selecting tests, critical reviews in the *Mental Measurements Yearbook* and in appropriate journals should be consulted in addition to test catalogues, and test specimens *always* should be examined before selection. Results of testing are confidential data but they should be communicated quickly and fully to the professional persons who are to use them. The cost of a testing program is no more than a fraction of one per cent of the total cost of instruction.

School or system-wide testing programs are used in connection with a number of school functions or purposes. Among them are guidance, ability grouping, and general instructional facilitation. In guidance, results of vocational interest tests should be considered suggestive only, and individual testing should be discussed with parents before use. Ability grouping should be based

on a composite of test results, school marks, and teacher reports, and not on a single test. Grouping is warranted for specific subjects only, not in general, and even then it leaves a wide range of differences within each section still to be accommodated. Knowledge of scores on intelligence and achievement tests enables teachers to plan materials, content, and method more adequately for their classes. It is considered inappropriate, if not invalid, to use standardized test results as a basis for marks or for promotion or to use them in an administrative evaluation of teacher effectiveness.

A testing program needs to be directed by a single person having administrative status. Teachers and principals should participate in decisions which affect them and concerning which they are qualified to judge. The director of the program should serve as a consultant to teachers in the administration and use of tests as well as strictly as a psychometrician and should co-ordinate whatever in-service training program is necessary for insuring efficiency in the program.

EXERCISES

1. Confer with the director of a testing program in a school district about his program. Find out the tests used, their frequency of administration, referral procedures, to what uses test results are put, special problems, etc.
2. Confer with several teachers about their activity in connection with the schools' testing program. Ask them to show you exactly how they use test results.
3. Devise an ability grouping plan for a secondary school. Indicate the courses for which there is to be grouping, the basis for the grouping in each case, the tests to be used, and how mistakes in pupil assignment are to be detected and corrected.
4. Plan a testing program for an elementary or secondary school or both. Determine what things are to be measured and how often. After an inspection of catalogues and test reviews, order specimen sets of the tests you think might be used. From these select the tests to be used in the program.
5. Sit with a school counselor when he discusses vocational plans with several students. Write a critique of his use of test data, academic records, etc.

BIBLIOGRAPHY

1. Bloom, B. S., and Heyns, I. De V., "Development and Application of Tests of Educational Achievement," *Review of Educational Research*, 26:71-88, February, 1956.
2. California State Department of Education, "Evaluating Pupil Progress," *Bulletin No. 6*, 1952.
3. Clark, W. W., "Evaluating School Achievement in Basic Skills in Relation to Mental Ability," *Journal of Educational Research*, 46:179-191, November, 1952.
4. Englehart, M. D., "Testing and Use of Test Results," *Review of Educational Research*, 26:5-13, February, 1956.

5. George, C. E., and Kingston, A. J., "The Stability of Interest Scores of College Freshmen," *Journal of Educational Psychology*, 46:243-246, April, 1955.
6. Hall, E. C., "The Proper Use of Test Results," *Elementary School Journal*, 54:450-455, April, 1954.
7. Jordan, A. M., *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953.
8. King, J. E., Jr., *Using Tests in the Modern Secondary School*, *NASS Bulletin*, 32:5-92, October, 1948.
9. Lass, A. H., and Wrightstone, J. W., "Evaluation in the Secondary Schools of New York City," *High Points*, 28:11-12, March, 1946.
10. Lefever, D. W., "Teacher's Role in Evaluating Achievement," *Education*, 71:203-209, December, 1950.
11. Leverenz, A. S., "New Developments in Evaluating Achievement in the Public Schools of Los Angeles," *Education*, 71:237-244, December, 1950.
12. Ludlow, H. G., "Trends and Issues in Standardized Testing," *Journal of Educational Research*, 47:279-284, 1953.
13. ———, "Trends in Public School Testing," *University of Michigan School of Education Bulletin*, 25:81-84, March, 1954.
14. Michaelis, J. A., "Current Practices in Evaluation in City Schools Systems," *Educational and Psychological Measurement*, 9:15-22, 1949.
15. ———, and Howard, C., "Current Practices in Evaluation in City School Systems in California," *Journal of Educational Research*, 43:250-260, December, 1949.
16. ———, "Evaluation in California County School Systems," *California Journal of Elementary Education*, 17:12-20, August, 1948.
17. Northby, A. S., "Secondary Education—III. Organization," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 1168-1169.
18. Otto, J. H., "Elementary Education—III. Organization and Administration," in *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: The Macmillan Company, 1950, pp. 377-378.
19. Parker, D. H., "Planning a Program of Standardized Testing," *High School Journal*, 36:107-112, January, 1953.
20. Patterson, C. H., "Predicting Success in Trade and Vocational School Courses: Review of the Literature," *Educational and Psychological Measurement*, 16:352-400, 1956.
21. Remmers, H. H., and Gage, N. L., *Educational Measurement and Evaluation*. New York: Harper Bros., 1943, Chaps. 10 and 20.
22. Rinsland, H. D., "Form for Briefing and Evaluating Standardized Tests," *Journal of Educational Research*, 42:371-375, January, 1949.
23. Ross, C. C., and Stanley, Julian C., *Measurement in Today's Schools*. New York: Prentice-Hall, Inc., 1954 (revised). Part IV.
24. Snodgrass, Florence, "Unreliability of Group Test Profiles," *Journal of Educational Psychology*, 45:129-142, March, 1954.
25. Thorndike, R. L., "Development and Applications of Tests of Special Aptitudes," *Review of Educational Research*, 26:14-25, February, 1956.
26. Tiegs, E. W., *Tests and Measurements in the Improvement of Learning*. Boston: Houghton Mifflin Co., 1939, pp. 262-285.

27. Traxler, Arthur E., *How to Use Cumulative Records*. Chicago: Science Research Associates, 1947.
28. Traxler, A. E., and others, *Introduction to Testing and the Use of Test Results in Public Schools*. New York: Harper Bros., 1953.
29. U.S. Office of Education. *State Testing and Evaluation Programs* (proc.). Circ. No. 320, 1951.
30. Young, P. A., "Use of Tests in Pupil Guidance," *National Association of Secondary School Principals Bulletin*, 32:73-84, March, 1948.

APPENDIX

APPENDIX A

GLOSSARY

- Ability tests.* Tests that purport to measure an individual's over-all facility in doing given things. Often a distinction is attempted between that facility which results from heredity and that which results from learning. In such cases, *ability tests* are usually applied to the "native" aspect and *achievement tests* to the learned aspect.
- Absurdity items.* Either statements or pictures that contain an element which is incongruent, inconsistent, contradictory, invalid, etc. The testee is required to pick out such absurd element. Used in intelligence tests and in measurement of critical thinking and reasoning ability.
Example: The rabbit chased the dog around the yard.
- Achievement tests.* Tests that purport to measure an individual's performance or competence relative to a given subject, usually a subject taught in the schools. Achievement tests are concerned with learned outcomes (generally knowledge and/or understanding) rather than "native" capacity or ability to learn the subject.
- Admission tests.* Tests or other measuring devices used to determine the eligibility of students for admission to schools or special curriculums. Ordinarily the province of colleges and the armed services.
Example: College Entrance Board Examinations.
Synonyms: *Selection tests* and *screening tests*.
- Age equivalents.* A method of expressing scores on standardized tests. The raw score typical of pupils of different ages is determined and then any pupil's raw score may be converted to the age to which it pertains. Usually given in years and months.
Example: Mental age = 12-6; reading age = 10-4.
- Age norms.* The typical scores made on a standardized test by pupils of different ages. Usually expressed in tabular or graphic form, raw scores being related to the correlative age in years and months.
Synonyms: *Age tables*, *age charts*, and *age conversion tables*.
- Alternate forms.* Standardized tests sometimes are issued in two forms containing different items, which yield scores relating to the same dimensions and have the same significance so far as age, percentile, standard score, etc. equivalents are concerned. The purpose of this is to permit retesting without undue practice effect. The *less used* form is called the alternate form.
Example: Stanford Revision of the Binet, Form M.
Synonyms: *Equivalent form* and *comparable form*.

Analogy items. A type of verbal or graphic item frequently used in tests of intelligence to measure reasoning ability, specifically the facility for generalization. They consist of two parts, one of which expresses a relationship or comparison and the other of which requires that the same sort of relationship or comparison be established among other elements.

Example: Pig is to bacon as steer is to: tree, fish, steak, shoes, corned beef.

Analysis of variance. The total variation of given measures for a group of individuals may be accounted for by variations among the individuals with respect to other factors. For example, the variation in IQ of a group of students may be attributable to their variations in age, cultural background, socioeconomic level, etc. The analysis of variance technique tests the statistical significance of the effect of each factor. That is, it tells the extent to which chance variation might have produced the same effect.

Anecdotal reports or ratings. Appraisals based on observations recorded in the words of the observer. Sometimes special forms called anecdotal records are used for the writing.

Appraisal. A general term meaning to determine and express the status of anything. Often used synonymously with measurement but usually connotes less precision in results. Sometimes used synonymously with evaluation, but tends to imply less judgment.

Appreciation tests. Instruments designed to measure attitudes and judgments relative to given subjects: usually art, music, and literature.

Approximation. A concept basic to measurement, which asserts that any given measure of a phenomenon can only "approximate" a "true" measure of its actual status. Unreliable procedures produce very gross approximations and more reliable ones produce the closer approximations. In standardized tests the degree of approximation involved is indicated by the standard error of score for the test in question.

Example: The IQ of 102 found by administration of a Binet test to a pupil only approximates the pupil's "theoretically true" IQ. This might be 100, 105, or even 93, etc. The Standard Error of Score of Binet IQ's is 4 to 5 for the middle ranges of intelligence.

Aptitude tests. Tests or other measuring instruments (usually standardized and commercially published) which purport to predict the ease with which an individual will learn a given thing or the degree of success he is likely to have in a given activity.

Example: *Stenquist Mechanical Aptitude Tests, Seashore Measures of Musical Talent.*

Arbitrary origin. See *assumed mean*.

Arithmetic mean. ($\bar{X}$) The sum of a group of scores divided by the number of scores in the group. Used as a representative score or a measure of central tendency.

Synonyms: *Average, mean.*

Arrangement items. A type of guided response question or item which presents the elements of a graphic, mechanical or verbal pattern and requires that the testee put them in proper array. Used fairly extensively in intelligence, personality, and mechanical aptitude testing.

Example: Make a sentence out of these words:

me letter the to brought postman a

Assumed mean. The point in a frequency distribution arbitrarily chosen to be the point from which the deviation of other measures is figured and from which a correction for the true mean is computed. Usually the mid-point of a central interval in a grouped distribution.

Synonyms: *Guessed mean, arbitrary origin.*

Attitude tests. Free or guided response tests which purport to measure the feelings of individuals toward certain things. Usually gauged are the valence or direction of feelings, the referents of feelings, and sometimes the intensity of feelings.

Example: Tests of attitude toward war, school, minority groups, etc.

Average. See *arithmetic mean.*

Average deviation. The sum of the absolute value of deviations (d) from the mean in a frequency distribution divided by the number of measures in the distribution. An infrequently used measure of dispersion for a group of measures.

$$\text{Average deviation} = \frac{\sum |d|}{N}$$

Bar graph. Any graphic presentation that uses bars of various lengths to symbolize differences in quantity, size, amount, etc.

Basal age. A convention of the Stanford Revision of the Binet Intelligence Scale.

The mental age designation of the last group of tests a child answers correctly in entirety. Contrasted with top or maximum age, which is the mental age designation of the last group of tests in which a child passes any test.

Battery. A lengthy standardized achievement test with separate and independent parts for each of several school subjects or skills; or a group of tests published by the same firm, applicable to the same grades, and standardized on the same population, hence producing comparable percentile or standard scores. Battery sometimes may mean any group of tests administered together for a given purpose.

Bimodal. A distribution of measures, particularly test scores, with two foci of central tendency rather than one. A superficial indication of bimodality is the presence of two modes separated by scores or score intervals whose frequency is appreciably less than that of the modes. Bimodality in a distribution can be suggestive of several attributes of the group or of the test or other measuring procedure in use. It often indicates that the group which is bimodal involves two subgroups having important mean differences as to age, mentality, reading ability, nationality, etc.

Biserial r . A correlation coefficient computed for two variables, one of which is a continuous normal distribution and the other is dichotomized (split in half). It is assumed that the underlying dichotomized variable is a continuous normal distribution.

Example: To find the correlation between honesty or dishonesty (a dichotomized variable) and intelligence test scores (a continuous variable) a biserial r would be used.

Case. A generalized term for any entity in a group that is separately measured or counted.

Example: *A pupil* in an eleventh-grade class, *one guinea pig* in a nutritional experiment involving 100 guinea pigs, *one teaching situation* in a study involving a comparison among many teaching situations.

Synonyms: *Individual, occurrence, entry.*

Ceiling. The highest degree of skill, knowledge, or any other dimension that a given test can measure. A perfect score on a test is its ceiling and pupils who achieve perfect scores may be said to have "hit the ceiling." A valid test must have a ceiling which exceeds that of the greatest degree of skill, etc., to be encountered in the group to be measured.

Central tendency. In a distribution of scores or other measures, the point or interval at which a plurality or majority of scores tends to cluster. Unless there is such a clustering, the distribution has no central tendency.

Chance factor. In any guided response, the correct option for any item *may be guessed* as well as known. The extent to which chance may be the cause of a correct response rather than knowledge or other rational determinant is a function of the number of options and the number of those that are correct for any item. *Chance factor* means that proportion of the maximum possible score on a guided response test which can be attributed to chance, or the odds against guessing, expressed either as a ratio or as a percentage.

Example: In a true-false test the chance factor is one/two, or 50 per cent. In a five-option multiple-choice test the chance factor is one/five, or 20 per cent.

Check lists. A device used in observation to direct attention to factors to be observed and sometimes to provide space for recording ratings or comments relative to them.

Chi-square (χ^2). The sum of the squared discrepancies between observed (O) and expected (E) frequencies, each divided by the expected frequency.

$$\chi^2 = \frac{\sum(O - E)^2}{E}$$

This statistic is generally useful in testing agreement with a priori frequencies, and in particular for testing goodness of fit, group differences, change, and independence.

Example: Chi square may be used to check sex differences in passing or failing a test item.

C.A. (chronological age). A child's age expressed in years and months. Used in reckoning the intelligence quotient and any other index involving a comparison between skill or knowledge and age.

Example: C.A. = 12-6 means the child's age is twelve years, six months.

Classification. One of four basic forms of measurement (types of measurement symbols). Involves the establishment of categories (classifications), the designation of symbols for the categories, and then the assignment of the symbols to phenomena according to the category to which they belong.

Examples: Blood typing; draft classifications; A, B, C, D, F as course marks.

Coefficient. A special name applied to certain ratios or proportionality constants.

Example: See *coefficient of correlation*.

Coefficient of correlation (r). Basically, it is a measure of the degree of closeness with which the variation of one variable is associated with variation of another variable. In this text it is shown that the square of the coefficient of correlation is equal to the ratio of the explained variance to the total variance.

It is important to note that in computing a correlation coefficient between two variables it is necessary to assume that the underlying distribution for each variable is a continuous normal distribution.

Example: Coefficient of correlation between intelligence and school marks is about .50.

- Completion items.* Test questions which require that one or more missing parts of a statement be filled in or completed. Classified as a guided response (or objective) item but may involve some interpretation in scoring.
Synonym: *Fill-in items*.
- Comprehensive tests.* Any tests that cover a wide range of subject matter. The term is used primarily at the college level to refer to achievement tests covering entire academic areas, e.g., economics, biology, education, etc.
- Confidence interval.* An interval used to estimate the value of a true score or a true mean in which a certain amount of confidence is placed.
Example: the chances are 95 out of 100 that a student's true IQ is between 92 and 108. The interval 92-108 is called the 95 per cent confidence interval.
See *level of confidence* and *level of significance*.
- Construct.* A verbal, mathematic, or graphic exposition offered as an explanation for natural phenomena.
Example: Atomic theory, Freudian psychiatric theory.
Synonyms: *Model, theory, hypothesis, explanation* (at some times).
- Contingency coefficient.* A measure of the degree of association or correlation between two variables whose variations are expressed in terms of categories. Chi square (χ^2) is used in its computation.
Example: A contingency coefficient would be used to indicate the correlation between A's, B's, C's, and D's received in a mathematics course, and A's, B's, C's, and D's received in a physics course.
- Controlled observation.* Observation of behavior in which those under observation are subjected to prearranged stimuli or in which the timing, focus, and recording of observations are highly systematized.
- Correction.* Statistically, a numerical quantity added to or subtracted from an estimate in order to obtain a true amount or at least a better estimate.
- Correction for attenuation.* A correction applied ordinarily to correlation coefficients that have been reduced or attenuated by variable errors of measurement. A correlation coefficient, then, which has been corrected for attenuation, is an estimate of what the correlation would be if based upon perfect and errorless measurements.
- Correction for guessing.* Any of several systematic ways of penalizing wrong answers more than omitted answers in true-false and multiple-choice tests. Such correction is based on the assumption that all students will try to guess many answers unless deterred from guessing by the knowledge that they will be penalized for it. It is intended to minimize the chance variation in scores.
- Correction formulas.* Formulas used in correcting raw scores of tests for guessing (see page 117). The effectiveness of these formulas is questionable.
- Correlation.* The tendency of the variation of one variable to be accompanied by the variation of another variable. A cause-effect relation is not necessarily inferred between the two variables.
- Covert dimensions.* Conditions, elements, or properties attributed to unobservable aspects of behavior, thinking, attitudes, drives, etc.
Example: Imagination, intensity of attitudes, and drive strength.
Roughly synonymous with *Inferred dimensions*.
- Criterion.* In general, anything with which a measuring procedure is compared in determining its validity. Specifically, a measuring procedure for a given phenomenon for which exemplary validity is claimed or assumed and with

which other similar procedures are asked to have high positive correlations. Example: The Stanford Revision of the Binet for intelligence, the Rorschach for personality analysis.

Synonym: *Standard* (in some contexts).

CR (Critical ratio). The quotient of the difference between two statistics divided by the standard of error of this difference. The ratio usually is interpreted to mean standard deviation units in a distribution of differences between the statistics that could be produced by chance; hence is an indication of the significance of the difference.

Example: Difference between percentage of students expressing liking for mathematics and those expressing dislike might be 32 per cent; the standard error of this difference might be 12 per cent; the Critical Ratio then would be $30/12 = 2.5$ and such a difference should be found on a chance basis only about 1 out of 100 times.

Culture-free tests. Intelligence tests have been criticized on the ground that familiarity with the content and values of middle-class Anglo-American culture affects the scores of pupils. Culture-free tests are those that purport not to involve the content or values of a given culture.

Cumulative frequency. A column in a conventional tabulation of scores or other measures that shows the frequency of scores up to and including any given interval.

Cumulative record. Any form used by a succession of teachers and/or counselors in recording data of importance to the academic progress and guidance of pupils. Intelligence, reading, and achievement test scores usually are included as well as subject grades, observations on deviate behavior, health information, and comments on social adjustment.

Curve fitting. A name given to the statistical methods for determining the equations of straight lines or curves that best fit the plotted points of a graph or scatter diagram.

Curvilinear relationship. Applies to the situation in which the plotted points of a graph of two variables approximate a curve rather than a straight line.

Example: Age as against height.

Decile. Any one of nine percentile points in a distribution of scores that divide the distribution into ten equal parts. The first decile is the 10th percentile, the second decile is the 20th percentile, and so on.

Derived score. A test score that has been converted to an index of rank, scale position, or classification, as distinct from a raw score, which is the number of correct responses or the immediate numerical weight given the test.

Example: Percentile rank, standard scores, mental age.

Synonym: *Converted score*.

Description. In this context, an informal type of measurement expression used to indicate the status of phenomena in which ordinary language is used. Descriptions may include scale, rank, and classification symbols. Associated with observation procedures and the appraisal of citizenship, study habits, social adjustment, etc.

Deviation. In general, departure from a given condition. In particular, the numerical difference between a test score or other measure of an individual and a given point of reference, usually the mean of a group of test scores or other measures.

- Deviation IQ.* An intelligence quotient determined by converting a raw test score to a standard score on a scale that has 100 as the mean and approximately 15 or 16 as its standard deviation. It is opposed to the more traditional *ratio IQ*, which compares mental age with chronological age. For children, deviation IQ's and ratio IQ's have essentially the same significance and may be compared.
- Diagnostic tests.* Tests designed to reveal points of strength and weakness in a pupil's skill or knowledge in a given subject. They are characterized by part scores and/or a profile. Some intelligence and personality tests are considered diagnostic in that they provide analytic scores.
- Digit span.* A convention of intelligence testing which refers to how long a series of numbers an individual can recall, having heard them spoken at second intervals. Used as an index of memory.
- Dimensions.* A collective term for properties, aspects, attributes, qualities, etc., of phenomena subject to measurement. Anything with respect to which we measure a phenomenon.
Example: *Height* of a pupil, *accuracy* of spelling, *rate* of reading.
Synonym: *Variable*.
- Discriminating power.* The characteristic of a test item to distinguish between two or more groups of people, usually those who have great knowledge of a given subject and those who have little, or those who manifest a given "trait" and those who do not.
- Distractor.* Any option in a multiple-choice or matching item that is incorrect.
- Distribution.* A table or graph showing the scores or other measures found for a group, so arranged that the number who have a given score or who fall within a given range of scores is apparent.
Synonyms: *Frequency distribution*, *frequency tabulation*, *distribution table*.
- Educational age.* When age norms are determined for a student in specific subjects such as reading, arithmetic, social studies, science, etc., the average of these age norms is called his educational age. The index is construed to mean that a pupil's level of academic achievement is comparable to that of the average pupil of the age given.
Example: John's educational age is twelve years, three months.
Synonymous with *Achievement age*, and may be converted into grade placement.
- Entrance examinations, tests.* Tests or other measuring procedures used in selective admission to schools and/or curriculums. Employed primarily by colleges and professional schools.
Example: College entrance board examinations, graduate record examinations, American Council on Education psychological examination.
Synonym: *Selection, admission tests or examinations*.
- Equivalent form.* Either of two forms of a measuring instrument, particularly a standardized test, parallel in content, difficulty, and norms, but different as to items.
Synonyms: *Alternate form*, *comparable form*.
- Essay tests, items, questions.* A type of free response test or item in which an extended written answer is to be given to a discursive question. Commonly used in secondary schools and particularly in colleges.
Synonym: "*Blue book*" examinations.

Evaluation. In this context defined to be the process of assigning symbols to phenomena, which symbols signify the worth of the phenomena relative to some scheme of value.

Evaluative standard. That with reference to which value is judged. In particular, a scale, hierarchy, or series of gradations or levels describing the variations of value in the status of a phenomenon and the value assigned to given gradations or levels.

Examination. A general term, synonymous with test in most cases, which denotes any procedure used for human measurement.

Examiner. A person who administers an examination or test.

Synonyms: *Tester, test administrator, test proctor.*

Extrapolation. Inferring measures outside the range of known measures on the assumption that any observed pattern in known measures will continue.

Example: It is known that the mean score of eighth graders on a given test is 80; ninth graders, 85; tenth graders, 90; and eleventh graders, 95. By extrapolation, the mean for twelfth graders is 100.

Factor. Anything treated as an entity which is known or presumed to be a significant part or aspect of a complex phenomenon. For example, in intelligence, the factor of memory; in handwriting, the factor of slant.

Synonyms: *Dimension, variable.* In "factor analysis" factor has special meaning and refers to that which underlies or causes positive correlation between scores on different tests or parts of one test.

Factor analysis. A term applied to methods based on the intercorrelations of several tests that attempt to account for the interrelationships among the tests in terms of a few underlying factors. Factor analysis has aided in the identification of basic components of intelligence, aptitudes, and personality through a study of the interrelationships of already prepared tests in these areas.

Fill-in items. A type of guided response item or question in which sentences are presented having one or more missing words. The testee is required to fill in the absent term(s).

Example: Tom Sawyer floated down the _____ River in the book called _____.

Floor. The least degree of knowledge, skill, intelligence, etc., that a test is capable of measuring. If any pupil has a raw score of zero (no questions answered correctly), his knowledge, etc., is below the "floor" of the test and hence the test cannot be used to measure him. A valid test must have a floor considerably below that of the least knowing, skilled, or intelligent person to whom it is to be applied.

Foil. A wrong or incorrect option in a multiple-choice item.

Example: For this question the underlined responses are foils.

The President of the United States in 1918 was *Harding*, *Hughes*, *Wilson*, *Roosevelt*, *Taft*.

Synonyms: *Distractor, incorrect option.*

Form of measurement. A type of number or other symbol used to characterize a phenomenon with respect to some dimension. Forms in use are scale numbers, rank numbers, classification numbers or symbols, and descriptive words or phrases.

Example: 12'6" is a *scale* form, 12th percentile is a *rank* form, *B* student is a *classification* form, and "makes more errors in punctuation than in spelling" is a *descriptive* form.

Free response tests and items. Tests and items or questions thereof in which a person may respond in his own words or with his own self-devised actions and in which the possibilities of response are many and the length of the response more than a word, phrase, or gesture.

Example: Explain the construction of a pleated skirt.

Synonyms: *Essay question* and, to some extent, "*subjective*" tests and/or items.

Frequency. Refers in statistics to the number of times a score is repeated or to the number of scores appearing in a given interval.

Frequency distribution. Consists of a sequence of score intervals and the frequency or number of scores falling in each interval.

Example: See page 133.

Frequency polygon. A line graph of a particular frequency distribution. The horizontal axis contains a scale for the score intervals and on the vertical axis is a scale of frequency per interval. The frequency polygon is then a series of straight lines connecting points at the middle of each interval representing the total frequency in each interval.

Example: See page 137.

Grades. Has two common meanings: 1. The 12 to 15 basic and sequential yearly subdivisions of elementary and secondary schools; 2. evaluative symbols assigned to pupil work, tests, or achievement to indicate its value. In the latter sense, *grades* are synonymous with *marks*. *A, B, C, D,* and *F* are the most frequently used letter grades or marks.

Grade equivalent. An index of achievement in which a pupil's ability or skill as measured by a standardized test is denoted by naming the grade for which the ability is typical.

Synonym: *Grade placement.*

Graph or graphic. Any visual representation of amount or quantity as related to other variables. Common types are histograms, frequency polygons, bar graphs, line graphs, and pie graphs.

Graphic rating scale. A line with numbers or a series of numbers that symbolize the range of variation in some behavioral dimension under or by which are written brief descriptions of several different degrees of the dimension.

Group evaluation. Any discursive activity in which a group analyzes and evaluates its activity, progress, or accomplishment. Used particularly in unit instruction and by committees or groups that pay attention to the tenets of "group dynamics."

Group tests. Tests administrable to a group of persons at one time. Contrasted with individual tests, administrable to only one person at a time.

Guidance folder. A folder used as a repository for standard tests, health records, teacher testimony, and other documents bearing on the ability, achievement, and adjustment of a school pupil. The folder may itself be a printed form for the recording of subject grades, test results, and other data.

Guided-response tests and items. A collective term for tests and test questions in which the possibility of response is strictly limited and the significance of any response often is predetermined.

Example: True-false tests and items, multiple-choice tests, completion tests, etc.

Roughly synonymous with *Objective tests and items*.

Histogram. A vertical bar graph of a frequency distribution. At each interval on the horizontal axis a vertical bar is erected whose scaled height represents the total frequency in the interval.

Example: See page 136.

Index. In behavioral measurement, any number or other symbol that signifies or indicates status with respect to some dimension.

Example: A given IQ is an index of intelligence, high school marks are indexes of college aptitude, a given coefficient of correlation between halves of a test is an index of the test's reliability.

Individual tests. Tests administrable to only one person at a time. Contrasted with *group tests*, see above.

Inferred dimension. In this text a property or quality of a phenomenon not itself observable but imputed or *inferred* to a phenomenon.

Example: Knowledge, reasoning ability, introversion, etc.

Usually synonymous with: *Covert and intangible dimensions*.

Ink blots. The vague and meaningless blots employed in a famous personality test, the Rorschach.

Instrument. Any device used to facilitate measurement. In education it is usually a piece of paper bearing printing or typing to which an individual is required to respond.

IQ (intelligence quotient). The most common index of a person's intelligence or brightness relative to what is normal for his age. IQ of 100 means average intelligence and higher IQ's indicate more, and lower IQ's less intelligence. For children it may be interpreted as a ratio between mental age and chronological age. For adults it may mean only their relative position among a general population of adults.

Intelligence tests, testing. Instruments and procedures, and their use, employed to measure intelligence, mentality, or general ability.

Synonyms: *Mental tests, tests of general ability*.

Intercorrelation. A term applied to each of the correlations among a group of tests. Intercorrelations are usually displayed in tables showing the correlation of each test with each of the other tests. Intercorrelations are then used to show the extent of interrelationships among a certain group of tests.

Interpolation. Estimating a value between two known points, linear relationship assumed.

Example: If a pupil's test score is 94 and a table shows that a score of 90 equals a mental age of ten years, and 102 indicates a mental age of ten years, six months, the pupil's estimated mental age would be ten years and two months.

Inventories. A name for a type of test that attempts to classify a pupil's interests, attitudes, values, etc.

Example: *Kuder Preference Vocational Interest Inventory, Bernreuter Personality Inventory*.

Interval, i. An arbitrary portion of a range or continuum of scores, usually designated by a lower boundary score and an upper boundary score.

Example: 23-27 is an interval in a score range of 16 to 75.

Item. A question or direction in a test that requires an answer from the person examined.

Synonyms: *Question, test question, test item.*

Item analysis. The process by which the relative difficulty and discriminating ability are determined for items in a test.

Item difficulty. The percentage of individuals who get the item correct.

Example: An item of 50 per cent difficulty would be answered correctly by half of those who responded to the item. An item of 85 per cent difficulty would be answered correctly by 85 per cent of those responding to it and hence would be a less difficult item.

Key. The correct answers for a test or a basis for interpreting answers. May be simply a test that has the correct answers marked, a stencil, or a coded sheet for an electrical scoring machine.

Labeling items. A type of guided response test item which asks the individual to name the designated parts of a drawing, map, or chart.

Level of confidence. A statistical term used to indicate the degree of confidence we may place upon an interval estimate. Level of confidence is usually indicated by a probability percentage, such as being 95 per cent sure or being 99 per cent sure.

Example: On the basis of a sample it is estimated that the true mean of a population lies somewhere between 64 and 67 and this interval was computed with a 95 per cent level of confidence. The chances, then, of the true mean being contained in the interval 64-67 are 95 out of 100. See *level of significance and confidence interval*.

Level of significance. A statistical term used to indicate the amount of confidence in whether or not the difference between two means, two percentages or other comparable measures is statistically significant (not due to chance).

Example: It is reported that the difference in mean scores of boys and girls is statistically significant at the .01 level. This means that the chances are only 1 out of 100 that the difference is not significant. See *confidence interval and level of significance*.

Linear relationship. If the plotted points of a graph of two variables closely approximate a straight line, then a linear relationship is said to exist between the two variables.

Machine scoring. The process of scoring a test by means of an electronic scanning device. It necessitates use of a special response sheet where answers are made by marking spaces or numbers with a pencil whose lead is an electrical conductor.

Man-to-man rating scales. A type of rating scale for recording appraisals of behavior in which the person rated is compared with a verbal description of a person whom he is thought to resemble.

Marks. The letter symbols (in some cases numbers) used to evaluate pupil achievement in school subjects and on tests and products related to the subjects. The most common marks are A, B, C, D, F.

Often synonymous with: *Grades.*

Matching items. A type of guided response item that involves two columns of associated items. The response consists of pairing the items. Used to test knowledge of wars-dates, authors-books, animals-species, relationships, etc.

Matrix. A tabulation of friendship or popularity preferences among a group.

Mean. See *arithmetic mean*.

Measurement. The assignment of a symbol, often a number, so as to characterize the status of a phenomenon relative to some dimension, usually by indicating its scale position, its rank, or its classification re this dimension.

Example: Finding a road is *ten miles* long, that a pupil ranks *third* in spelling, that a pupil's interests are *mechanical*.

Roughly synonymous with: *Appraisal*.

Measurement error. Error in a statistical sense is a matter of chance factors and not of mistakes made. Consequently, measurement error is the difference between the true value and observed value of a measurement due only to whatever chance factors may be present and not due to a mistake of the measurer. The standard error is used to estimate the amount of such measurement errors.

Median. The score or point that divides a distribution of scores into two equal groups, with half of the scores falling above and half below. Used as a representative score or a measure of central tendency.

Synonyms: *Middle score*, *50th percentile*.

MA (mental age). The age group whose average mental development is most like that of the child in question.

Mental tests, testing. Tests or testing procedures that purport to measure differences in intelligence or general mental ability.

Example: *Wechsler Intelligence Scale for Children*, *California Test of Mental Maturity*.

Synonyms: *Intelligence tests*, *general ability tests*, etc.

Mid-point of interval. The halfway point between the actual boundaries of an interval. It is found by adding half the size of the interval to the actual lower boundary or subtracting from the actual upper boundary.

Example: The mid-point of the interval 33-36 is 34½.

Mode. The score or measure that occurs most frequently in a distribution.

Model. A system of verbal, mathematical and/or graphic symbols, purporting to explain any phenomenon and that may be the basis for testing, experimentation, prediction, instruction, etc.

Example: Euclidean geometry, quantum theory of light, general and special factors of intelligence, Darwinian theory of evolution.

Synonyms: *Theory*, *construct*, *hypothesis*, etc.

Multiple-choice items. Test items or questions widely used in standardized tests, which permit the testee to choose an answer from among several options only one of which is correct or each of which has a special significance.

Example: The author of *The Adventures of Tom Sawyer* is

1. Grey, 2. Cooper, 3. Twain, 4. London, 5. Melville.

Synonym: *Multiple-option items*.

Multiple correlation. A multiple correlation is computed when more than two variables are involved. Suppose we have three variables, *x*, *y*, and *z*, and suppose variables *y* and *z* are combined to estimate variable *x*. The correlation between *x* and the combination of *y* and *z* which best estimates *x* is called a multiple correlation.

Example: High school marks and college entrance exam scores may be combined to predict first-year college grades. Hence a correlation between first-year college grades and the combination of high school marks and entrance

exam scores that predicts college grades is called a multiple correlation.

N. In statistics, *N* stands for the number of measures in a sample or in a distribution.

Negative discrimination. Students may be separated into two groups, those who who made high scores on a given test and those who made low scores. The proportion of either group who were correct on a given item on the test may then be determined. If a greater proportion of the group making low scores got the item correct than those who got high scores, the item is said to show *negative discrimination*. *This is considered evidence of invalidity in the item.*

Nonverbal tests and items. Tests and items so developed that variation in responses to them is not importantly affected by differences among people in verbal skill.

Example: Form board tests, geometric and mechanical puzzles, picture recall and discrimination items, manipulatory tests and items.

Synonym: *Performance tests and items*.

Normal curve. Refers to the normal probability curve, which is a theoretical distribution of probability described precisely by a mathematical equation. The curve has a distinct bell-shaped appearance. Some distributions of observed variables closely approximate the shape of a normal probability curve. When this occurs, then certain probability statements may be made in connection with these variables.

Example: If a distribution of reading scores which approximates a normal curve has a mean of 62 and standard deviation of 10, then we could say that the chances are only 16 out of a hundred for a student to get a score of 72 or higher on the test.

Normalizing. The process of readjusting test scores in a distribution to make them conform to the normal distribution, the assumption being that the true distribution of the trait being measured would be normal if there were a satisfactory measure of it.

Norming. The process of establishing equivalence between raw scores on a standardized test and age, grade placement, percentile rank, and standard scores in the population on which the test is being standardized.

Synonyms: *Standardizing, standardization*.

Norms. Statistics based upon a standardization group or a group that is purported to be representative of a much larger population. These norms are thus assumed to be representative of large groups such as all fifth-grade children or all twelve-year-olds. Grade, age, percentile, and standard score norms are the most common form.

Objective tests. Measuring instruments that are amenable to mechanical, electronic, or other scoring method little dependent on the interpretation or judgment of the scorer.

Example: True-false and multiple-choice tests.

Synonym: *Guided response tests*.

Observation. The most widely used and usually most crude method of behavioral measurement. Involves direct perception of the dimensions of the phenomenon being measured. With appropriate attentional, perceptual, and recording aids, observation can be a highly reliable procedure.

Example: Watching a pupil study to determine how effective are his study habits.

Observation schedules. Sheets of paper with captions, directions, etc., used to direct observation and often as a means of recording observations.

Opinionnaire. Any instrument designed to elicit the written opinion of respondents on given questions. Subsumed by the more general term, questionnaire, which deals with questions of fact as well as opinion.

Options. The response variants from among which an individual may select his answer to a test question.

Example: Columbus sailed to the West Indies in

(a) 1470, (b) 1492, (c) 1592, (d) 1607.

The four dates listed are options.

Synonyms: *Choices, alternatives.*

Overt dimensions. Aspects or properties of behavior subject to measurement (dimensions) that are directly observable (overt).

Example: Rate of talking, slant of handwriting, strength of grip.

Synonyms: *Observable dimensions, objects of direct measurement, tangible properties.*

Parent-teacher conferences. A discussion between parent and teacher regarding the progress of a pupil. Used with increasing frequency in addition to or as substitute for a report card.

Partial correlation. The correlation between two variables where the influence of another variable or other variables has been eliminated.

Percentile. A standard index of relative position or rank in a distribution of scores which consists of that percentage of scores lying below any given score or point.

Example: The 65th percentile is that score or point below which lie 65 per cent of the scores.

Performance test. Any test or other procedure of measurement that is not importantly influenced by an individual's verbal skill and that purports to measure some nonverbal dimension of intelligence or achievement.

Example: Form boards, recall of pictures, spatial perception tests, etc.

Somewhat synonymous with *nonverbal tests*.

Personality tests. Any tests that purport to measure the differences among individuals as to the nature and dynamics of their goals, fears, identifications, needs, drives, adjustment mechanisms, etc.

Pie graph. A graphic procedure for displaying proportions of a total amount by marking off sectors of a circle that are proportionately equal to the relative size of the amounts in question.

Population. Used in an abstract sense in measurement and statistics to indicate any given group of things (pupils, schools, teachers, experimental animals, etc.). Often means the totality of any group in question as opposed to a sample of this group.

Synonymous with *universe*.

Positive discrimination. In achievement testing, the characteristic of a test item (usually guided response) to be answered correctly more often by students making high scores on a test in question than by those making low scores. Test items must show positive discrimination if they are valid for the dimension to which they relate. See: *Negative discrimination*.

Power tests. Specifically, measuring instruments whose items show increasing dif-

ficuity when arranged in serial order. In general, measuring instruments that are not timed and are designed to measure the extensiveness or depth of an individual's knowledge or skill.

Practice effect. It is known that a performance of any task affects a reperformance of that task, usually in the direction of improvement. "Practice effect" is the term for the significance of such reperformance when the same test is administered to the same individual more than once.

Pretest. Any measuring instrument (usually an achievement test) administered prior to a period of instruction, an experiment, or other circumstance of interest. As a rule, pretests are used to establish the initial status of pupils so that the amount of their learning may be judged from the results of a later retest.

Probability. As applied to behavioral measurement, the concept that any measure or statistic is somewhat subject to chance variation. Hence it deviates from some theoretically "true" measure. Such deviation is commonly called error and its probable extent can be determined and stated mathematically.

Probable error. A statistic derived from the standard error used to establish an interval estimate for which there is 50 per cent confidence. For normal distributions, the probable error is equal to .6745 time the standard error.

Example: If the mean of a sample is 47 and its probable error is 4, then the chances are 50 out of 100 that the true mean lies between 43 and 51.

Product analysis. A basic procedure of educational evaluation in which the things that pupils produce in the course of instruction are appraised in appropriate ways and given scores or ratings.

Example: Compositions, outlines, drawings, wood work, etc.

Product-moment formula. A widely used formula for determining the correlation coefficient. Let z_x be the standard score for variable x and let z_y be the standard score for variable y . If the pairs of z_x 's and z_y 's for each individual are multiplied, then added for all individuals and divided by the number of cases, the result is the product-moment formula for the correlation coefficient.

$$\text{Correlation coefficient } (r) = \frac{\sum z_x z_y}{N}$$

Thus, according to the product-moment formula, the correlation coefficient is the mean of the set of products of standard scores for the two variables.

Profile. An analytic graphic presentation of a pupil's scores on a test battery, scores on parts of a given test, marks in several school subjects, ratings on several personality variables, etc.

Prognosis and prognostic. The matter of predicting the future accomplishment of individuals on the basis of systematic measurement, usually guided response testing.

Synonym: *Prediction and predictive.*

Prognostic tests. Any measuring instruments that serve effectively in predicting the future accomplishment of individuals.

Example: Tests of musical aptitude, intelligence tests, college entrance examinations.

Synonyms: *Predictive tests, aptitude tests.*

Projective techniques, methods, tests. Free-response procedures for behavioral measurement that present an ambiguous, vague, or unstructured stimulation

to the testee. His response (oral, written, or graphic) then is presumed to reveal or *project* his personality. Used extensively in diagnosing mental disorders.

Example: The *Rorschach Ink Blot Test*, the *Murray Pictures*.

Psychometrist. A person who engages professionally in psychometry.

Psychometry. The body of concepts, procedures, principles, devices, etc., connected with the measurement of behavioral or psychological phenomena.

Quartile. Any one of three percentile points that divide a distribution of scores into four equal groups or quarters.

Q (quartile deviation). A common index of variation or dispersion in a distribution of scores. It is the distance between the upper and lower quartiles divided by

two. It is sometimes called the semi-interquartile range. $Q = \frac{Q_3 - Q_1}{2}$

Example: If, in a distribution of scores having a range of from 32 to 98,

$Q_3 = 71$ and $Q_1 = 53$, $Q = 9$ or $\frac{71 - 53}{2} = 9$.

Q_1 (*first quartile*). The lower quartile, or 25th percentile, which marks off the lower 25 per cent of a distribution of scores.

Q_3 (*third quartile*). The upper quartile, or 75th percentile, which marks off the upper 25 per cent of a distribution of scores.

Questionnaire. Any of a number of paper forms used to get information from a person.

Quintile. Any one of four percentile points that divide a distribution of scores into five equal groups; every twentieth percentile. The first quintile is the 20th percentile, the second quintile is the 40th percentile, and so on.

Random sample. A sample selected from a larger population in such a manner that every individual or object in the population had an equal chance of being chosen in the sample.

Range. The difference between the highest and lowest scores in a given distribution of scores.

Example: If the highest score in a distribution were 74, and the lowest were 30, the range would be $R = 74 - 30 = 44$.

Ranking, rank numbers. The process of ordering the constituents of a group in terms of some dimension. Rank numbers indicate the relative position of the constituents.

Example: Finding and stating the rank of students as to skill in tennis.

Rate tests. Tests which measure the speed at which individuals can perform certain activities, as reading, typing, shorthand.

Synonym: *Speed tests*.

Rating. A direct appraisal of a dimension in terms of some descriptive scale or verbal classification scheme.

Example: Rating the citizenship of pupils from poor to excellent; rating school plants as unsatisfactory, satisfactory, excellent.

Rating scale. A verbal or pictorial device used to facilitate and to record ratings as described above.

Example: See page 56.

Raw score. The first quantitative untreated result obtained in scoring a test.

Example: 85 right on a spelling test.

Readiness tests. Any of a variety of measuring instruments and procedures used to see if pupils are "ready" to attempt to learn a given subject, most commonly reading.

Reading age (RA). A type of "norm" score derived from standardized tests used to indicate a pupil's reading ability in terms of age equivalents. Reading age means the age group whose average performance is most like the performance of the child in question. RA may be established only in terms of a *given* standardized test.

Example: Mary's reading age is 12-3, or twelve years and three months.

See: *Norms*.

Reading grade. A type of "norm" score derived from standardized tests that states a pupil's ability to read in terms of grade equivalents. Reading grade means the school grade whose average performance is most like that of the pupil in question. By interpolation, the reading grade may be fractional. As with reading age, reading grade refers only to a given standardized test.

Example: John's reading grade is 6-7, meaning like the average child who has been in the sixth grade for seven months.

See: *Norms*.

Reliability. An essential characteristic of a measuring instrument having to do with consistency and dependability. An instrument is reliable to the extent to which it yields the same results when reapplied to the same dimension of the same phenomenon whose status has not changed. Reliability is indicated by high coefficients of reliability and by low standard errors of score.

Report cards. A means of communicating pupils' progress and achievement of educational objectives.

See: Appendix C.

Representative sample. A sample of a population so drawn as to reflect accurately all essential and pertinent characteristics of the population.

See: *Random sample* and *sampling*.

Representative score. Any single score used to represent a group of scores. Ordinarily this may be the arithmetic mean, median, or mode.

Retest. A test readministered at the end of a period of instruction or other activity, the results of which are to be compared with an earlier administration of the test.

Rho (ρ). The rank-difference measure of correlation. Individuals are assigned ranks with respect to each of two variables, and for each individual the difference (d) in rank is determined. These differences are squared and summed for all cases and substitution is made in the following formula.

$$\rho \text{ (rho)} = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

Sampling. Selecting a portion of a population on which to base estimates of the population.

Example: Political polls, test standardization samples, ore sampling.

See: *Random sampling*, *representative sampling*.

Sampling error. Errors due to the chance factors in random sampling. Sampling error is usually estimated statistically by the standard error.

Example: The difference between the mean of a sample of a population and the mean of the entire population.

See: *Measurement error*, *standard error*, *confidence interval*.

Scaling, scale numbers. Measurement in terms of defined and precise units that represent given amounts or degrees of some dimension. Scale numbers indicate the number of units and hence the amount or degree of the dimension. Scale numbers must refer to a fixed point of reference, usually a zero.

Example: Measuring height with a ruler, speed with a watch, and intelligence in terms of deviation IQ's.

Scattergram. A two-dimensional chart used as a basis for computing a correlation coefficient for two variables, x and y . The chart consists of a horizontal scale for the x -variable and a vertical scale for the y -variable. Pairs of x and y values are plotted as points or tallies on the chart. The scatter of points gives a visual picture of the relationship between the x and y variables.

Example: See page 176.

Scoring. The process of assigning a score (usually a number or letter symbol) to a test or pupil product. For a test, this is often done by comparing a paper with the key, marking the questions answered correctly, and adding up the total.

Scoring stencil. A mechanical device for scoring tests.

See: Page 117.

Screening tests. Procedures used by colleges, armed forces, and other agencies to eliminate persons who fail to exhibit certain minimum qualifications.

See: *Admissions tests*.

Self-evaluation. Any of many concepts and procedures concerned with an individual observing and judging his own performance, achievement, or adjustment.

Short answer items. Types of guided response questions to which a pupil may respond with one or a few words, the significance of which is exactly predetermined.

Sigma, σ . In statistics, the small Greek letter σ (sigma) is a symbol for the standard deviation of a theoretical distribution. More particularly, and with subscripts, it symbolizes the theoretical distribution of means, correlation coefficients, etc., and also stands for standard error.

Significance of differences of scores, etc. The degree to which chance factors probably did not produce the difference observed.

See: *Confidence interval, probability, level of confidence, measurement error*.

Skewed. The characteristic of some distributions of measures to cluster not at the middle of the range but toward either extreme.

See: Page 149.

Sociometry. The concepts and procedures having to do with the measurement of personal or feeling relationships among members of groups.

Split-half reliability. A coefficient of reliability based upon splitting a test into supposedly equivalent halves and correlating the scores on one half of the test with paired scores on the other half of the test.

Standard. See: *Evaluative standard*.

Standard deviation. An index of variation in a group of measures. It represents the square root of the mean of the squared deviations of the individual measures or

$$s = \sqrt{\frac{\sum d^2}{N}}$$

Where d is the deviation of a score from the mean.

Standard error. The estimated standard deviation of an imaginary normal distribution of repeated samplings. The standard error is used as the basis for computing confidence intervals or interval estimates of true values.

Example: A student's test score is 78 and the standard error is 4. If the student were retested infinitely, his test scores would form a normal distribution and the standard deviation of this imaginary curve is called the standard error. Consequently, the chances are 68 out of 100 that the student's true score lies in the interval 74–82, or the 95 per cent confidence interval would be 70–86 for the true score.

Standard score (z score). A general term referring to any of a number of scores that indicate how many standard deviations a measurement is above or below the mean. It is found by determining the difference between the raw score (x) and the mean ($\bar{x}$) and dividing by the standard deviation (s).

$$z = \frac{x - \bar{x}}{s}$$

In many standardized tests, the z scores are converted to positive and whole numbers by conversion formulas. For example, a z score of -1.5 , which means that the student in question is one and one-half deviations below the mean, might be converted to a test "standard score" of 35 by the following conversion formula:

$$(z \text{ score of } -1.5)10 + 50 = 35$$

Synonyms: *Sigma score*, *z score*.

Standardization population. The individuals who were administered a given test prior to publication and on whose scores are based the norm scores for the test.

Standardized tests. Tests, usually published, which have been preadministered to a population of known characteristics and yield scores in terms of this population. This population is selected so as to be a representative sample of the total population for which the test is designed.

Stanine. Any one of nine intervals on a scale of standard scores. The stanine (abbreviation for standard-nine) scale spans the normal curve in nine intervals of size equal to one half of a standard score. The stanine intervals have values from 1 to 9 and the middle interval, 5, extends from standard score $-\frac{1}{4}$ to $+\frac{1}{4}$.

See: Appendix D.

Statistic(s). Any derived quantity obtained from a set of raw scores or measures.

Examples: N , mean, standard deviation, median, mode, quartile deviation, correlation coefficient.

Statistical significance. A difference is said to be statistically significant when there is only a slight probability that the difference was due to chance variation. Formerly, the term was used only when the size of a difference exceeded its standard error by threefold.

See: *Confidence interval*, *confidence level*.

Strip key. A simple type of key for scoring a guided response test in which correct answers are shown on a strip of paper rather than on a whole test. The strip usually is the answer column of the test.

Subjective tests. Tests scored without a key and on the basis of the examiner's interpretation and judgment. Usually they involve long written responses.

Synonym: *Essay test*.

Survey tests. Measuring instruments or procedures designed to measure broad areas of knowledge or ability in terms of one or a few general dimensions. Opposed to diagnostic tests, analytic tests, and profile tests.

T-score. A special conversion of *z* scores or standard scores in order to eliminate plus and minus signs and decimals. *T* score conversions are particularly applied to *z* scores that are medians of normal curve intervals representing categorical ratings like good, fair, excellent, and so on. The median *z* score is multiplied by 10 and added to 50.

$$T \text{ score} = 10z + 50.$$

Tests. Any of a great number of procedures in which individuals respond to common stimulations and react in comparable ways and which yield a measure of the individuals with respect to one or more dimensions.

Synonyms: *Examinations, instruments, quizzes*.

Test plan. In standardized test construction, the rationale of the test, including what is to be measured, how, and how expressed.

Test protocols. Directions for interpreting, classifying, and/or rating responses to individually administered intelligence and personality tests.

*Tetrachoric *r*.* A special correlation coefficient that is computed for two variables with the variation of both expressed as dichotomies.

Example: The correlation between good or poor students and their passing or failing a test question. The assumption is made that the underlying distribution of both variables is a normal distribution.

Thought questions. A loose term meaning all sorts of test items that involve reasoning as well as memory and that usually require extended answers.

Example: What is the relationship between Shakespeare's plays and Elizabethan politics?

Timed tests. Tests for which students have an allotted time to respond.

Top. The highest degree of any dimension that a given test can measure. Usually a valid test needs to have more top than any individual to whom it is to be administered. This is indicated by an absence of perfect scores.

Synonym: *Test ceiling*.

Traits. The elemental attributes of personality according to certain theories and test plans. Comparable to "factors" in intelligence.

Example: Honesty, sympathy, etc. Aggressiveness, dominance, etc.

True-false items. A form of guided response item in which pupils indicate whether they think statements are correct or incorrect.

True score. The theoretical measure that a valid and reliable test would produce if no chance variables affected it. Statistically, it is the mean of the distribution of an infinite number of measurements made of the same thing.

See: *Measurement error, standard error*.

Universe. See: *Population*.

Untimed tests. Tests which all students are expected and allowed to finish.

See: *Timed tests*.

Validation. The process of establishing on the basis of empirical data the validity of a test, usually a standardized one, by comparing its results with one or

more criterions. Typically involves, as a minimum, item analysis, correlation of results with other test scores, analysis of distributions of scores, and determination of reliability.

Validity. The essential characteristic of a measuring procedure or instrument *actually* to measure the phenomenon and dimensions that it purports to measure.

Variable. In measurement, a term for any significant property, aspect, condition, etc., of a phenomenon whose variance affects variance in the phenomenon. For example, variation in the school achievement of pupils is affected by the variables of intelligence, age, motivation, instruction, etc.

See: *Dimension*.

Vocational interest tests. Tests designed to determine the specific vocations or occupational areas in which a person may be interested.

z score. See: *Standard score*.

APPENDIX B

ANNOTATED BIBLIOGRAPHY OF SELECTED PUBLISHED TESTS

Explanation:

1. Tests listed are considered to be of average or better validity and usefulness and to represent the variety of available tests. Detailed analyses of these and all other tests are to be found in Buros' *Mental Measurements Yearbooks*, in reviews in periodicals, in publishers' catalogues, and in manuals for the tests.
2. An effort has been made to cover all age and grade levels and most subjects of importance in the schools.
3. The inclusion of a test does not indicate its unqualified endorsement, nor should an omission be considered to reflect on the test omitted.
4. Except where there are clear and specific factors bearing on a test's validity, validity is not discussed and may be presumed to be consistent with prevalent standards for standardized tests.
5. Reliability coefficients are stated where possible and, unless stated otherwise, are derived from split-half coefficients of correlation. In some cases data bearing on reliability are vague but there still is an indication of reliability equivalent to that of comparable tests. Here the statement "satisfactory" has been used in lieu of a coefficient. Where no data are available, this is stated.
6. The comments column contains an analysis of subtests or dimensions covered by the test, types of score, and any particularly important features. Detailed reviews are beyond the scope of this bibliography.
7. Cost of tests is to be found in publishers' catalogues. Cost is mentioned only where it is unusually high (as for certain individual tests kits). Separate answer sheets for either manual or machine scoring are available for most tests listed. Catalogues will specify.
8. This list should be used only to indicate possibilities and to order specimen tests and manuals, *not* to adopt tests without further inspection and evaluation.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
<i>General Achievement Batteries</i>					
California Achievement Test Batteries 1933-51	California Test Bureau	Four Batteries	Timed		Age, grade, and percentile norms for total and subtest scores.
		Primary Grades 1-3	90 min.	.91-.98 on subtests	Each level contains:
		Elementary Grades 4-6	120 min.	.77-.94 on subtests	Vocabulary, reading comprehension, arithmetic reasoning, and fundamentals, mechanics of English, and spelling.
		Intermediate Grades 7-9	150 min.	.83-.94 on subtests	A significant feature of the tests is their provision of comparable measures of achievement, grades 1 through 14.
		Advanced Grades 9-14	150 min.	.81-.93 on subtests	
Cooperative School and College Ability Tests (SCAT) 1956	Educational Testing Service	Grades 10-14	70-min. or two 45-min. periods	Kuder-Richardson Verbal .93	Percentile norms by grade. Careful preparation and validation procedures. .50-.70 correlations are reported with school grades. The tests are essentially reading, vocabulary, and arithmetic tests. Little knowledge of the content of school subjects is involved, nor is much reasoning or critical thinking.
		Level 1: 13-14	Timed	Quantitative .91	
		Level 2: 10-12		Total .95	It is intended to extend the tests downward into the middle elementary grades. They are designed ultimately to replace the ACE as an

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(SCAT continued)					
Essential High school Content Battery 1951-52	World Book	Grade 10 through college	5 class periods —1 of 45 min. and 4 of 40 min. Timed	Split-half for whole test .95 Alternate form .67-.92 for separate tests	instrument for educational placement. Shows the standard error of scores on profile, an excellent innovation. Percentile norms by grade for total score and subtest: Mathematics Science Social Studies English, according to (a) students in general (b) academic students (c) nonacademic students
Iowa Every-Pupil Tests of Basic Skills 1940-47	Houghton Mifflin	Elementary Battery, Grades 3-5	Timed	Subtests below .85	Grade placement norms. Longer than most achievement batteries. Claim diagnostic value, but subtest reliabilities may be too low. Manual should be checked.
A. Reading			50 min.	Total scores higher	
B. Work-study skills			55 min.		
C. Usage and spelling			60 min.		
D. Arithmetic			60 min.		

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(Every-pupil Tests continued)					
A. Reading		Advanced Battery, Grades 5-9	85 min.		
B. Work-study skills			90 min.		
C. Usage and spelling			70 min.		
D. Arithmetic			80 min.		
Iowa Tests of Educational Development 1942-48	Science Research Associates	Grades 8.5-13.5	8 hrs. for total battery Timed	.81-.94 for subtests	Percentile norms by grade on a profile sheet for each subtest: 1. Social concepts. 2. Natural sciences. 3. Usage. 4. Quantitative thinking (general mathematics). 5. Social studies reading. 6. Natural science reading. 7. Literary comprehension and appreciation. 8. Vocabulary. 9. Using sources of information. S.R.A. scores whole battery edition. Separate editions for each subtest for local scoring. Useful in determining educational level of transfer students and those with irregular school experience.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Metropolitan Achievement Tests 1931-1950	World Book	5 batteries covering Grades 1-8 Primary I, Grades 1-2 Primary II, Grades 2-3 Elementary, Grades 3-5 Interim, Grades 5-7 Advanced, Grades 7-9	Timed 60 min. 100 min. 2½ hrs. 4 hrs. 4 hrs.	.79-.97 for a given grade	Age, grade, and percentile norms for total and subtests as follows: Vocabulary, phrases, and numbers. Reading, vocabulary, arithmetic, and spelling. Reading, vocabulary, arithmetic, usage, and spelling. Reading, vocabulary, arithmetic, English, literature, history, geography, science, and spelling. Reading, vocabulary, arithmetic, English, literature, geography, science, and spelling. Subject tests published separately as well as in booklet.
Stanford Achievement Test (Revised) 1953	World Book	4 batteries covering elementary grades	Timed	Subtest reliabilities range from .71-.95	Content may be somewhat dated because of effort to retain pattern and form of earlier editions. Age norms, grade norms, and percentile norms by grade for total scores and subtests, as follows:

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(Stanford continued)					
		Primary Grades 1.9-3.5	80 min.		Paragraph comprehension, vocabulary, spelling, and arithmetic.
		Elementary Grades 3-4.9	135 min.		Vocabulary, paragraph, comprehension, usage, arithmetic, and spelling.
		Intermediate Grades 5-6	215 min.		Paragraph meaning, social studies, science, usage, arithmetic, spelling, and study skills.
		Advanced Grades 7-9	215 min.		As for intermediate.
					Subtests also published separately.
					A revision of the widely used earlier test.

Graves Design Judgment Test, 1948	Psychological Corporation	Grades 7-16	<i>Art</i> Nontimed, 20-30 min.	.81-.93	Percentile norms based on small samples. Perception of unity, dominance, variety, balance, continuity, symmetry, proportion, rhythm. Designs are abstract and non-objective to minimize effect of content. Carefully constructed, difficulty of validation on universal criteria, effect of mental set on scores unknown.
					\$1.50-\$1.75 per test.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Meier Art Tests: I. Art Judgment, Revised, 1942	Bureau of Educational Research and Service, State University of Iowa	Grades 7-12	Non timed, 45-60 min.	.70-.84	Percentile and quartile norms. Based upon responses to two-dimensional expression, linear and pattern qualities in composition, adapted from works of reputable artists. 85¢-\$1.00 per test.
Business Relations and Occupations; Achievement Examinations for Secondary Schools, 1951	Educational Test Bureau	High School	<i>Business</i> 65 min. Timed	No data	Norms in terms of median and quartile deviation. Prepared by a committee of high school teachers for use in Minnesota.
National Business Entrance Tests, 1949	Joint Committee of United Business Assoc. and Natl. Office Managers Assoc.	Grades 12-16	Timed	No data	Administered only at test centers, which may be established in any community.

Form 1391-
Business
Fundamentals
and General
Information

75-85 min.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Form 1392—Bookkeeping Test			120–130 min.		
Form 1393—General Office Clerical Test (including filing)			120–130 min.		
Form 1394—Machine Calculation Test			120–130 min.		
Form 1395—Stenographic test (long and short form)			Long, 115–125 min.; short, 60–70 min.		
Form 1396—Typewriting Test (long and short form)			Long, 120–130 min.; short, 60–70 min.		
S.R.A. Dictation Skills, 1947. Two parts: Speed, Accuracy.	Science Research Associates	Grades 9–12 Adults	40 min. Untimed.	.80	Dictation is taken by any stenographic system. Test consists of determining missing words. \$18.75 per set of records, \$1.90 per 25 answer pads. Norms available.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
<i>English</i>					
(See also General Achievement Batteries)					
Brown-Larken Listening Comprehension Test, 1953	World Book	Grades 9-13	50 min. Untimed	.84-.90	Percentile norms by grade. Administered orally to a class as a whole.
Cooperative English Test, 1950-51	Educational Testing Service	2 levels: Lower, Grades 7-12; Higher, Grade 11 through College	40 min. each test. Timed	Satisfactory	Percentile norms for total and subtests. Test A: Mechanics of Expression. Test B: Effectiveness of Expression. Test C: Reading comprehension (speed, level, and vocabulary). As with any nonwriting test, validity regarding actual composition ability may be questioned.
Cooperative Literary Comprehension and Appreciation Test, 1935-51	Educational Testing Service	Grade 10 through College	40 min. Timed	Satisfactory for subject	Percentile norms. Deals with interpretation and discrimination rather than knowledge.
Gates-Russell Spelling Diagnosis Test, 1937	Bureau of Publications Teachers College, Columbia Univ.	Grades 2-6	Untimed	No data	Yields 9 scores as follows: Oral spelling, word pronunciation, letter sounds, spelling one-syllable words, spelling two-syllable words, word reversals, spelling attack, auditory dis-

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(Gates-Russell Continued)					
Michigan Vocabulary Profile Test, 1937-49	World Book	Secondary through adult	50 min. Untimed	.74-.94 for subtests (equivalent form)	crimination, visual-auditory, kinesthetic, and combined study methods. Individually administered. Percentile norms for high school, college, and certain occupational groups. Eight subtests cover human relations, commerce, government, physical sciences, biological sciences, mathematics, fine arts, and sports. Should not be used as sole basis for estimating aptitude for given occupations or studies.
Test of English Usage, 1950	California Test Bureau	High School and College	110 min. Timed	Part scores .85 Total .90 or more (Kuder-Richardson)	Percentile norms by grades for: I. Punctuation and capitalization. II. Grammar. III. Sentence and paragraph structure. Stresses verb forms in II and paragraph organization in III.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
<i>Foreign Language</i>					
Cooperative Foreign Language Tests: French, Latin, Spanish, 1942-51	Educational Testing Service	High School and College 2 levels for each language: Beginning and Advanced	40 min. Timed	Satisfactory	Percentile norms. Reading, vocabulary, and grammar; for Latin includes knowledge of civilization as well. In view of usual variety of emphasis and instruction in courses too much significance should not be attached to norms.
Useful for test-retest procedure.					

Handwriting

Ayres Handwriting Scale, Gettysburg Edition, 1912-17	Educational Testing Service	Elementary Grades	Function of examiner	A sample of a child's handwriting is rated by assigning it the number (10, 20, . . . 80) of the scale sample it most resembles. Tabulations on the scales show the percentage of children at each grade level who are <i>likely</i> to have handwriting like given samples. Nondiagnostic, but best present means of measuring children's handwriting.
--	-----------------------------	-------------------	----------------------	---

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Kilander Health Knowledge Test, 1951	World Book	High School	<i>Health</i> 40 min. Untimed	.88	Percentile norms. Tests knowledge of health and safety practices and human physiology.
Minnesota Tests for Household Skills, 1952	Science Research Associates	High School and College	<i>Home Economics</i> 30-40 min. each test	.73-.90	Based essentially upon knowledge of best practices in household activities.
a. Child care b. Cleaning c. Laundering d. Foods (Forms A and B)					
State High School Tests for Indiana.	State High School Testing Service for Indiana, Purdue University		Timed	No data	Percentile norms. Highly factual in content. Designed especially for Indiana students and courses of study so norms should be applied with caution to other than Indiana schools.
a. Food Selection and Preparation, Form C, 1948		High School	60 min.		

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(State High continued)					
b. Planning for Family Food Needs, Form C, 1948		High School	60 min.		
c. Helping with Food in the Home, Form A, 1948		Grades 7-8	35 min.		
d. Helping with the House-keeping, Form A, 1945		Grades 7-8	35 min.		
e. Home Care of the Sick, Form B, 1948		High School	60 min.		
f. Housing the Family, Form B, 1948		High School	60 min.		

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Middleton Industrial Arts Test, 1951	Kansas State Teachers College of Emporia	High School	<i>Industrial Arts</i> 65 min. Timed	.86	Percentile norms. Comprehensive sampling of factual learning in woodworking, with some attention to mechanical drawing.
A.C.E. Psychological Examination, Current	Educational Testing Service	College, particularly Freshmen	<i>Intelligence</i> 65 min. Timed	No data given, but presumed high	Percentile scores only; revised yearly; current edition scored by publisher. Most widely used college general ability test; separate linguistic and quantitative scores; timed feature puts older students at a disadvantage.
Army General Classification Test (AGCT), (identical to Form 1A of Army edition), 1947-48	Science Research Associates	Grade 9 to adult	50 min. Timed	.95	Standard scores with a mean of 100 and a standard deviation of 20. Doubtful use in the schools but important to recognize because of millions of servicemen's experience with it.
California Test of Mental Maturity, 1936-51	California Test Bureau	5 levels, covering kindergarten through adult	Long and short forms. Time varies according to level. Up to 60 min. All levels timed.	.92 to .95 for total scores, less for part scores	Furnishes a "profile" of intellectual scores, both verbal and nonverbal factors. MA's, IQ's, and grade percentiles.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Gesell Developmental Schedules, 1925-45	Psychological Corporation	4 weeks to 6 years	20-40 min. Untimed	Depends upon examiner	Yields age norms. Individually administered. Observation of behavior. Test based on Gesell studies at Yale. (\$74.00 for complete materials.) Useful mostly for research.
Goodenough Intelligence Test, 1926	World Book	Age 4-10	10 min. Untimed	Depends upon examiner	Mental ages. The old and well-known draw-a-man test. Especially good for children with verbal disabilities.
Henmon-Nelson Tests of Mental Ability, 1931-50	Houghton Mifflin	Elementary and High School	35 min. Timed	Data insufficient but presumed low because of shortness of test	Percentiles as well as MA's and IQ's. Verbal emphasis, simple to score and administer, single score only.
Holzinger-Crowder Unit-Factor Tests, 1955	World Book	Grades 7-12	Two 45 min. working periods Timed	Alternate form .81-.95. Split-half corrected, .88-.95	Furnishes grade percentiles and a general scholastic aptitude index. Converts to an IQ. Four factor scores: verbal, spatial, numerical, and reasoning.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(Holzinger continued)					
Northwestern Intelligence Tests, 1949-51	Houghton Mifflin	13-36 weeks	20-30 min. Untimed	.84 for total age range	Correlations with academic subject grades around .50; with nonacademic, somewhat lower. Yields IQ's. Reflects Gesell's studies. Relation of infant IQ to childhood IQ open to question except in extremes.
Ohio State Psychological Test, 1919-50	Science Research Associates	Grade 9 to adult	120 min. Untimed	Form 21 $r = .93$	Percentile norms for grades 9 through 13. Particularly useful with older college students who do poorly on timed tests. Used as a predictor of college aptitude. Continually revised since first publication in 1919.
Pintner General Ability Tests: Non-Language Series, 1945	World Book	Grades 4-9	50 min. Timed	.87	Heavy emphasis on verbal ability. Yields MA's and IQ's. Minimum verbal component. May even be administered by pantomime. Subtests are figure dividing, reverse drawings, pattern synthesis, movement sequence, manikin and paper folding.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Revised Stanford-Binet Scale (Forms L & M), 1937	Houghton Mifflin	Age 2 and up	Approximately 1 hr. in normal situation. Untimed save for specific tests.	Depends upon examiner but high potential	Furnishes MA's and Ratio IQ's. The "criterion" of intelligence tests. Individually administered by qualified examiner only. Now dated and many items are culturally misfit.
S.R.A. Primary Mental Ability Tests, 1938-50	Science Research Associates	3 levels: Age 5-7	2 subtests are timed 60-80 min.	Satisfactory	\$16.00 for complete set of materials. MA's and Deviation IQ's for totals and subtest norms as follows: 5 subtests: verbal-meaning, space, perceptual speed, motor and quantitative—mental age and IQ equivalents 7 subtests: verbal meaning (reading and non-), space, reasoning (reading and non-), perception and number—IQ for readers, IQ for nonreaders. 5 subtests: verbal meaning, space, reasoning, number, word fluency: several ability quotients (like IQ) and percentile norms.
		Age 7-11	60 min.		
		Age 11-17	45 min.		Effect of reading ability is minimized.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(S.R.A. continued)					
Wechsler-Bellevue Intelligence Scale, 1939-47	Psychological Corporation	Age 10 through adult	40-60 min. Untimed	Equivalent to Binet. Depends upon examiner.	Based upon Thurstone's differentiation of specific mental factors. Some reviewers question the validity of the differentiated factors and scores relative to them. Furnishes deviation IQ, which in school age children is equivalent to Binet. The only rival to the Binet for individual testing. Many prefer for older youth; most prefer for adults. Permits personality analysis as well as intelligence. Individually administered by qualified examiner. \$16.50 per set of materials.
Wechsler Intelligence Scale for Children (WISC), 1949	Psychological Corporation	Age 5-15	40-60 min. Untimed	Depends upon examiner, but high potential	Deviation IQ's equivalent to Binet IQ's. Modification of Wechsler-Bellevue to accommodate to children. Not yet as well accepted as the Binet for school age children. Individually administered by qualified examiner only. \$22.00 per set.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Arithmetic Test (Fundamentals and Reasoning), Municipal Battery, National Achievement Tests, 1938-39	Acorn Publishing Company	Grades 3-6, 6-8	<i>Mathematics</i> 70 min. Timed	No data	Percentile norms according to IQ classifications. Covers computation, verbal problems, understanding of number relationship, ingenuity in item construction. No information on background of preparation or experimental results.
Cooperative Algebra Tests: Elementary Algebra through Quadratics, 1948-50	Educational Testing Service	High School	45 min. Timed	Satisfactory	Percentile norms. Adequate coverage of mechanical and manipulative skills, verbal problems including expressing symbolic relationships, and graphs.
Cooperative Intermediate Algebra Test: Quadratics and Beyond, 1948-50	Educational Testing Service	High School	45 min. Timed	Satisfactory	Percentile norms. Gives attention to functional understanding of algebra, includes interpretation of graphs and formulas and language of variation.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Functional Evaluation in Mathematics, 1952	Educational Testing Service		Timed	Satisfactory	Percentile norms for total and subtests. 1 & 4, Quantitative understanding. 2 & 5, Problem solving. 3 & 6, Basic computations.
Tests 1, 2, 3		Grades 4-6	30 min.		Attempts to reflect scope and emphasis of modern arithmetic program. Some unrealistic problems.
Tests 4, 5, 6		Grades 7-9	30 min.		
Lankton First Year Algebra Tests, 1951-52	World Book	Grades 9-13	50 min. Timed	Split half, .87 Alternate form, .81	Percentile norms. Emphasizes meaning and skill in the understanding and use of the language of algebra.
Metropolitan Achievement Tests (Arithmetic), 1947-49	World Book	Grades 3-4 Grades 5-6 Grades 7-9.5	Timed 75 min. 90 min. 90 min.	.87-.95	Percentile norms. Limited to computation and solving verbal problems. Careful construction.
Snader General Mathematics Test, 1951-52	World Book	Grades 9-13	50 min. Timed	.80-.84	Percentile norms. For a general mathematics course which includes arithmetic, intuitive geometry, simple algebra, and numerical trigonometry. One third of items on lower grades arithmetic, emphasis on mechanics and memory.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Stanford Achievement Test (Arithmetic), 3 levels: Primary, Intermediate, Advanced, 1940-46	World Book	3 levels: Grades 2-3 Grades 4-6 Grades 7-9	Timed 25-30 min. 50-55 min. 60-65 min.	Satisfactory	Percentile norms. Emphasizes computation and solving verbal problems. Would not be appropriate for "meaning" phases of the modern arithmetic program.
Seashore Measures of Musical Talent, Revised Edition, 1939	Psychological Corporation	Grades 5-8, adults	<i>Music</i> 60-70 min. Untimed	.62-.79 for subtests.	Decile norms for scores on discrimination in pitch, loudness, time, timbre, rhythm, and tonal memory. Carefully constructed, useful in uncovering talent.
Wing Standardized Tests of Musical Intelligence: A Test of Musical Ability on 10 Records, 1948	Sheffield City Training College, Sheffield, England	Ages 10 and over	60 min. Untimed	.91 for total subtests, .65-.86	\$12.00 per set of three recordings. Uses recordings to test chord analysis, pitch change, memory, rhythmic accent, harmony, intensity, phrasing. Contains same basic constituents of musical ability as Seashore Test. More useful with above average students in musical talent.
					Norms based on English schools. \$20.00 per set.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
<i>Personality-Character-Interests</i>					
(Standardization of tests is particularly difficult in this area and hence published tests are neither as adequate nor as much used as for intelligence and achievement. Group tests are particularly suspect.)					
Allport-Vernon Study of Values, 1931-51	Houghton Mifflin	College and adult	20 min. Untimed	.87-.92, test/retest	Six relative "value" scores which measure dominant interests in personality as to theoretical, economic, aesthetic, social, political, religious. Considerable evidence of relation between scores and selected groups assumed to have given values.
Behavior Preference Record, 1953	California Test Bureau	3 levels: Grades 4-6, Grades 7-9, Grades 9-12	30-45 min. Untimed	.77-.91 (only 275 cases)	Consists of problem situations followed by options of action and reasons therefor. Answers relate to knowledge of acceptance of ideals and practices of democracy. Separate scores for friendliness, co-operation, integrity.
California Test of Personality, Revised, 1953	California Test Bureau	Grades 4-8	50 min. Untimed	.94	Separate norms by sex. Percentile norms for personal and social adjustment. Useful for screening for serious maladjustment but not for personality classification. Items susceptible to insincere responses.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Minnesota Multiphasic Personality Inventory (Group Form), 1942-51	Psychological Corporation	College and adult	40-90 min. Untimed	.56-.91 according to scale	Thirteen scores showing relative deviation from a normality group in these personality dimensions: Hypochondriasis, depression, hysteria, psychopathic deviate, masculinity and femininity, paranoia, psychosthenia, schizophrenia, hypomania, and social introversion.
Kuder Preference Records (Vocational), Form C, 1948-51	Science Research Associates	Grade 9 to adult	40 min. Untimed	Average .90 for scoring categories	Any finding should be verified by interview, individual testing, and case study as needed. For use by psychologist or psychometrist only.
Mooney Problem Check List; 1950 revision, 1941-50	Psychological Corporation	Forms for Grades 7-9, 9-12, 13-16, adults	20-40 min. Untimed		Percentiles for eleven vocational areas shown on a profile sheet. Self-scoring. Widely used but results should be treated as suggestive, not definitive. No norm scores. Not a test but a self-reporting instrument. Items grouped by areas: Health and physical development, courtship, sex and marriage, home and family, etc. Use is suggestive and exploratory, not for measurement of problems.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Rorschach Ink Blot Test, 1921 to date	Grune-Stratton, Inc.	Age 3 and over	Variable	A function of the examiner	Most widely used clinical instrument for classifying and diagnosing personality disorder. Involves verbal response on part of testee to ink blots and interpretation by examiner according to protocol.
					Used only by qualified psychometrists and psychologists.
					\$19.50 for complete materials.
E. S. Bogardus Social Distance Scale, 1925-51	The author, 3518 University Ave., L.A.	Age 15 and over	25 min. Timed	No data	Not a test but a self-reporting device. An experimental instrument that has had an important effect on measurement of attitudes (see page 412).
					Scales for:
					a. Ethnic distance
					b. Occupational distance
					c. Religious distance
					d. Economic distance.
					More used in sociological studies of group attitudes and status than for measuring individuals.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Strong Vocational Interest Blanks, 1927-51	Stanford University Press	College and adult	40 min. Untimed	.88	Widely used vocational interest tests. Separate forms for men and women. Specific classification scores for 47 occupations (men) and 28 (women). Scored by the publisher for fee of \$1.10 to \$1.50 per blank. Hand and machine scoring blanks available at higher price than publisher-scored.
Thematic Apperception Test (T.A.T.), 1943	Harvard Press	Age 7 and up Different pictures according to age and sex	120 min. approximately Untimed	No data	Next to Rorschach, most used "projective" personality device. Individuals tell stories about vague pictures and these are interpreted according to test protocol in terms of manifestations of need (motivation) and press (environmental pressures). To be used only by psychometrists and clinical psychologists. \$6.00 for materials.

Reading

(See also *General Achievement Batteries*)

Durrell-Sullivan Reading Capacity and Achievement Tests, 1937-45	World Book	Two levels: Primary Grade 2.5-4.5 Intermediate Grade, 3-6	Untimed 40-45 min. 60-75 min.	.86-.95	Age and grade norms. Easier portions of both intermediate tests.
--	------------	---	---	---------	---

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(Durrell-Sullivan continued)					
Gates Basic Reading Tests, 1942	Bureau of Publications, Teachers College, Columbia University	Grades 3.5-8	Timed, varies according to grade. Up to 40 min. over-all	.80-.96 per subtest per grade	Test of capacity: understanding spoken words, understanding spoken discourse. Test of achievement: vocabulary, paragraph comprehension. The tests provide a basis for distinguishing between the slow learner and the normal child who has reading difficulty. Age and grade placement norms for: A. General significance. B. Predict outcomes. C. Understand directions. D. Note details. Appraises both speed and accuracy for each of the tasks indicated, hence somewhat diagnostic.
Gates Reading Diagnostic Tests, 1926-45	Bureau of Publications, Teachers College, Columbia University	Elementary	Untimed 60-90 min.	No data	21 norm scores relating to all aspects of reading disability. Individually administered. Used for retarded readers.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Lee-Clark Reading Readiness Test, 1951	California Test Bureau	Kindergarten—Grade 1	20 min. Timed	.93 on total, .83-.94 on subtests (N = 170)	Percentile norms for total scores and for subtests. 1. Visual discrimination. 2. Vocabulary and following instructions. 3. Identification of letters and words. Percentile norms. Entirely pictorial. Directions given orally but responses made by marking.
Metropolitan Readiness Tests, 1933-50	World Book	Kindergarten—Grade 1	65-75 min. Part non-timed	Alternate form .89	Percentile norms. Directed at auditory and visual discrimination and at rate of learning for words.
Murphy-Durrell Diagnostic Reading Readiness Test, 1949	World Book	Primary	60 min. for Tests 1 & 2 which are untimed. Test 3 is timed & administered at intervals during the school day.	.96 and .95 for Tests 1 & 2 No data for Test 3	Percentile norms. Includes numbers as well as just reading readiness, so generally predictive for first grade success.
S.R.A. Reading Record, 1947	Science Research Associates	Grades 8-13	40 min. Timed	Satisfactory if rigidly timed	Percentile norms by grade for total and subtests as follows: 1. Rate. 2. Comprehension. 3. Paragraph meaning.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
(S.R.A. continued)					
4. Reading a directory. 5. Interpretation of maps, charts, and graphs. 6. Advertisement reading. 7. Index usage. 8. Technical vocabulary. 9. Sentence meaning. 10. General vocabulary.					
Great emphasis on speed. Slight errors in timing may produce spurious results.					
Self-scoring					

Anderson Chemistry Test, 1951-52	World Book	Grades 11-13	<i>Science</i> Timed 50 min.	Split half .93; alternate form .87	Percentile norms. Emphasis upon understanding of principles, familiarity with laboratory work. Lacks computational questions, good coverage.
Cooperative Biology Test, 1947, 1948	Educational Testing Service	High School	45 min. Timed	Satisfactory	Percentile norms for high school biology classes. Adequate coverage of basic biological information, contains some application and interpretation of principles.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Cooperative General Science Test, 1947-50	Educational Testing Service	High School	40-45 min. Timed	Satisfactory	Percentile norms. Emphasizes recall of factual information. Lacks attention to methods of science. No manual.
Cooperative General Science Test for Grades 7-9, 1947-48	Educational Testing Service	Grades 7-9	80-85 min. Timed	Satisfactory	Percentile norms. Part I—General information. Part II—Largely terminology. Part III—Comprehension and interpretation. Part III is more valid from a curricular standpoint.
Cooperative Physics Test, 1947-49	Educational Testing Service	High School	50 min. Timed	Satisfactory	Percentile norms. Distribution of questions on mechanics, heat, light, sound, and electricity. Numerical problems. Primarily factual.
Dunning Physics Test, 1951-52	World Book	Grades 11-13	50 min. Timed	Split half .90	Percentile norms. Balanced coverage of mechanics, heat, sound, light, electricity. No emphasis on scientific method and application to everyday devices.
Nelson Biology Test, 1951-52	World Book	Grades 9-13	50 min. Timed	Split half .88; alternate form .77	Percentile norms. Emphasis upon application, interpretation, and problem solving. Excellent manual.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Read General Science Test, 1951-52	World Book	Grades 9-12	50 min. Timed	Split half .88; alternate form .85	Percentile norms. Excellent statistical procedures used in development of test. Items well constructed. Half of items devoted to facts and principles, other half to applications. Coverage reflects present general science programs.

Social Studies
(See also *General Achievement Batteries*)

California Tests in Social and Related Sciences (Advanced Battery) 1953-54	California Test Bureau	Grades 9-12	Timed 45 min. 45 min. 80 min. 170 min. total	Kuder-Richardson Formula .85-.97 .83-.98 .81-.92	Grade placement and percentile norms for total scores and subtests: 1. American History to 1876. 2. American History since 1876. 3. Physical and Biological Sciences. Administered as a battery or separately.
College Entrance Examination Achievement Test in Social Studies, 1937-51	Educational Testing Service	Grades 12-13	70 min. Timed	.90 or better	Norms for college Freshmen. Mostly American history, some world history. Includes interpretation of cartoons and weighing of arguments (available only in College Entrance Examination Admissions Testing Program). Strictly a "screening" test.

Title	Publisher	Age or Grade Range	Time	Reliability	Comments
Cooperative American History Test, 1932-51	Educational Testing Service	High School and College	45 min. Timed	Satisfactory	Percentile norms. Little attention to 1945 on.
Crary American History Test, 1951-52	World Book	Secondary	40 min. Timed	.87-.91	Good face validity and reviewers commend. Percentile norms.
Cummings World History Test, 1951-52	World Book	Secondary	40 min. Timed	.91-.94 (161 cases only)	Aspects of achievement covered: Factual, skill, interpretation, understanding, inference. Percentile norms.
Watson-Glaser Critical Thinking Appraisal, 1952	World Book	Grade 9 to adult	50 min. Untimed	.90-.95	Too little attention to events since World War II. All fifteen-year-olds and below have no recollection of other than post-World War II. Percentile norms for high school and college. Replaces an earlier test by the same authors.

PUBLISHERS' ADDRESSES

Acorn Publishing Company, Rockville Centre, N. Y.
Bureau of Educational Measurements, Kansas State Teachers College of Emporia,
Emporia, Kansas
Bureau of Educational Research and Service, State University of Iowa, Iowa City,
Iowa
Bureau of Publications, Teachers College, Columbia University, N. Y.
California Test Bureau, 5916 Hollywood Blvd., Los Angeles, California
Educational Test Bureau, 720 Washington Ave., S.E., Minneapolis, Minnesota
Educational Testing Service, Princeton, N. J.
Grune-Stratton, Inc., 381 Fourth Ave., New York, N. Y.
Harvard Press, 44 Francis Ave., Cambridge, Massachusetts
Houghton Mifflin Company, 2 Park Street, Boston, Massachusetts
Joint Committee of United Business Association and National Office Managers
Association, 132 West Chelton Ave., Philadelphia, Pennsylvania
Psychological Corporation, 522 Fifth Ave., New York, N. Y.
Science Research Associates, 57 West Grand Ave., Chicago, Illinois
Stanford University Press, Stanford, California
State High School Testing Service for Indiana, Purdue University, Lafayette,
Indiana
World Book Company, 313 Park Hill Ave., Yonkers, N. Y.

APPENDIX C

SAMPLE REPORT CARDS

Date of Placement	Check Level	READING ACHIEVEMENT by LEVELS	ATTENDANCE REPORT							
			Period	1	2	3	4	5	6	Tot.
		Review	Present							
		Lev. 1 - Readiness								
		Lev. 2 - Chart - Pre-Primer	Absent							
		Lev. 3 - Primer								
		Lev. 4 - First Reader	Tardy							
		Lev. 5 - Beginning Second								
		Lev. 6 - Advanced Second	Weight							
		Lev. 7 - Beginning Third								
		Lev. 8 - Advanced Third	Haight							

REMARKS

First Period.....

Second Period.....

Third Period.....

Fourth Period.....

Fifth Period.....

Sixth Period.....

1. ARITHMETIC.....
Is learning the meaning and use of numbers.....
Knows and uses number facts for his level.....

2. ART.....
Expresses himself freely.....

3. LANGUAGE.....
Expresses himself interestingly and clearly.....
Speaks distinctly.....
Is learning to speak correctly.....

4. MUSIC.....

5. PENMANSHIP.....
Writes legibly.....
Writes with reasonable speed.....

6. READING.....
Shows interest in stories and reading.....
Makes use of phonics.....
Reads with understanding.....
Reads well aloud.....

7. SPELLING.....
Is learning to spell new words.....
Spells correctly in composition.....

8. SOCIAL STUDIES.....
Takes responsibility for his share of the work.....
Tries to solve his own problems.....
Is increasing his knowledge of health.....
Is increasing his knowledge of science.....
Is increasing his understanding of his world.....

SOCIAL STUDIES PROJECT FOR.....

1st. Period.....

2nd. Period.....

3rd. Period.....

4th. Period.....

5th. Period.....

6th. Period.....

S-Satisfactory Progress I-Improving N-Needs Much Improvement
✓ Indicates Satisfactory Attainment

Kindergarten-Primary

GROWTH IN PERSONAL AND SOCIAL ATTITUDES AND HABITS (CITIZENSHIP)		GROWTH IN SCHOLARSHIP		PARENT AND TEACHER COMMENT	
This is a report on a number of attitudes and habits which are considered necessary to good citizenship. The school and the home share joint responsibility for their development. Consequently space has been provided for both school and home to check and to comment upon the same thing.		Class assignments to be completed at home are not recommended in the elementary grades. First school projects are limited to those work should be allowed to supplement their classroom work by outside preparation.		To Parents:	
It is the aim of the school to plan for its pupils happy, wholesome learning experiences. To benefit from class life each boy and girl should grow in habits of responsibility, self-discipline, and responsibility, and should develop attitudes of friendship, cheerfulness, and initiative. Continuous growth in attainment and performance should be evident as the pupil advances through the grades.		Borrow any book help that you can get for providing opportunities for hand work materials pertinent to the Area of "Citizenship" and for supplementing a wide variety of reading. In the elementary grades, specific instruction is best given in the school. Areas of Responsibility:		No one reporting method can give you the whole picture of your child's growth and development. The value of this report depends upon the attention you give it. Talk it over with your child. Your aim is to his classroom and arrange for a conference with his teacher.	
KEY: A-Attitude and Habit Are Developing Satisfactorily for Age C-Home Not Comment on Page 4 or 5		KEY: A-We have been able to demonstrate with Ability C-Home Not Comment on Page 4 or 5		In the following space teachers will inform you of any special interest or abilities your child has displayed. You will also find special commendation and constructive criticism written here. Part of the space will be reserved each time for your use. Let us know of the most valued activities of your boy or girl. Suggest ways in which school and home can supplement each other. All of us want to achieve the fullest development of your child's capabilities.	
REPORT PERIOD		REPORT PERIOD		FIRST REPORT PERIOD	
1 2 3 4 5 6		1 2 3 4		1 2 3 4	
CITIZENSHIP	Contributes to classroom activities	Participates in group planning	Responsible to school and community activities	Responsible to speech and action	Responsible to school and community activities
COURTESY	Respects personal habits	Respects personal habits	Respects personal habits	Respects personal habits	Respects personal habits
SELF-DISCIPLINE	Follows safety rules	Follows safety rules	Follows safety rules	Follows safety rules	Follows safety rules
SERVICE	Follows safety rules	Follows safety rules	Follows safety rules	Follows safety rules	Follows safety rules
WORK AND STUDY HABITS	Follows safety rules	Follows safety rules	Follows safety rules	Follows safety rules	Follows safety rules
FUNDAMENTALS OF LEARNING AND SOCIAL LIVING		FUNDAMENTALS OF LEARNING AND SOCIAL LIVING		FUNDAMENTALS OF LEARNING AND SOCIAL LIVING	
Reading		Reading		Reading	
Writing		Writing		Writing	
Arithmetic		Arithmetic		Arithmetic	
Social Studies (Geography, History, Civics, etc.)		Social Studies (Geography, History, Civics, etc.)		Social Studies (Geography, History, Civics, etc.)	
Health		Health		Health	
Physical Education		Physical Education		Physical Education	
Music		Music		Music	
Art		Art		Art	
Creative Arts		Creative Arts		Creative Arts	
Other Studies		Other Studies		Other Studies	

Elementary

Pupil _____

KEY TO MARKING
S—Growth is satisfactory.
NS—Is capable of doing better.

GROWTH IN SOCIAL BEHAVIOR

	Period	1	2	3
Enjoys playing with others				
Is willing to take turns				
Respects the property of others				
Is courteous to playmates				
Listens when other people are talking				
Uses soft voice indoors				

GROWTH IN HEALTH AND SAFETY BEHAVIOR

Keeps objects out of his mouth				
Uses handkerchief				
Uses fountains and lavatories properly				
Relaxes and rests willingly				
Carries equipment safely				
Follows school traffic rules				
Remembers his address				
Takes off and puts on own wraps				

GROWTH IN WORK BEHAVIOR

Listens to and follows directions				
Begins work quickly				
Asks for help when needed				
Works well with the group				
Takes care of school materials				

GROWTH IN OTHER AREAS

Informal reports will be made to parents on growth in the following areas:

Language

- Participates in conversation.
- Enters into dramatic play.
- Speaks plainly and with ease.
- Listens with interest to stories and poems.
- Enjoys looking at picture books.

Social Studies and Natural Science

- Is exploring his home and school environment.
- Is gaining understandings of group life.

Number

- Is becoming aware of number in daily living.

Music

- Sings with group.
- Dramatizes songs and instrumental selections.
- Enjoys quiet listening.

Art

- Enjoys beautiful things in nature and art.
- Expresses ideas in drawing, painting, and modeling.
- Experiments with many different materials and tools.

Play

- Enjoys big-muscle activities such as running, jumping, and climbing.
- Is gaining motor control, body power, and strength through appropriate play activities.
- Expresses ideas in simple body rhythms, games, and stunts.

TEACHER COMMENTS

First Period: _____

Second Period: _____

Third Period: _____

Kindergarten

ACHIEVEMENT	FIRST QUARTER	SECOND QUARTER	THIRD QUARTER	FOURTH QUARTER
SCHOOL SUBJECTS				
Reading				
Language				
Writing				
Arithmetic				
Spelling				

A check ☒ indicates pupil's progress in terms of test scores and teacher's judgment.

Progress is to be interpreted in terms of pupil's ability only, and not used as a basis of comparison with the achievement of other pupils.

ATTENDANCE	FIRST QUARTER	SECOND QUARTER	THIRD QUARTER	FOURTH QUARTER
PAID DATE ABSENT				
FIRST TRAIL				

A check mark indicates teacher judgment.

A child is one of the children below will be made in the table called to the attention of the parent. The parent should be advised to keep the child's progress in mind.

CHILD'S NAME _____

DATE _____

TEACHER'S NAME _____

SIGNATURE OF PARENT OR GUARDIAN:

First Quarter _____

Second Quarter _____

Third Quarter _____

Fourth Quarter _____

PARENT CONFERENCE RECORD BLANK

GRADE 7-12

DATE _____

NAME _____ GRADE _____

SCHOOL _____ TEACHER _____

ABILITY AND INTERESTS AND ACHIEVEMENTS OF THE CHILD:

CHILD'S MOST IMMEDIATE NEED (PHYSICAL, EDUCATIONAL, SOCIAL):

WHAT THE TEACHER PLANS FOR THE CHILD _____

WHAT THE PARENT PLANS FOR THE CHILD _____

Primary

Parent-Teacher Conference Blank

PUBLIC SCHOOLS	PUPIL'S REPORT CARD	JUNIOR HIGH SCHOOL		
RECORD OF _____		TEACHER _____		
GRADE _____	NUMBER OF DATE ABSENT <table border="1" style="display: inline-table; vertical-align: middle;"><tr><td style="width: 20px; height: 20px;"></td><td style="width: 20px; height: 20px;"></td></tr></table>			PRINCIPAL _____
DATE OF FIRST REPORT _____		DATE OF SECOND REPORT _____		

SUBJECT	ACHIEVEMENT IN SUBJECT	WORK HABITS	SHARING GROUP RESPONSIBILITY
	HIGH	HIGH	HIGH
	AVERAGE	AVERAGE	AVERAGE
	LOW	LOW	LOW

NEEDS IMPROVEMENT IN _____

PUPIL DOES BEST IN _____

SIGNATURE OF PARENT

RETAINED IN _____

(OVER) PROMOTED TO _____

FORM 411-117 USE 2-10-2004 11-31-11

Junior High

[illegible]

Junior High

[illegible]

Senior High

[illegible]

Senior High

REPORT OF UNSATISFACTORY WORK

Parent's Name

TO THE PARENT: *Address*

for the *week ending*

..... *has been doing unsatisfactory work in:*

SUBJECT	READER	BOOK

AN ANALYSIS OF THE DIFFICULTIES FOLLOWS:

	DIFFICULTY BELOW AVERAGE	BELOW AVERAGE	AVERAGE	Above AVERAGE	DISTINCTLY ABOVE AVERAGE
1. Interest in Subject					
2. Attendance and Punctuality					
3. Attention to Class Work					
4. Mastery of Facts					
5. Ability to Organize Material					
6. Ability to Recite					
7. Written Work					
8. Preparation of Daily Work					
9. Habits of Study					
10. Classroom Conduct					
11.					

For statement of work to be made up, see other side

Now—in addition to this report, grade reports are made to parents through students each succeeding nine weeks of the school year. You may examine carefully these reports and in conference with the student's teachers. In completing grades daily class work counts (three-fourths) and the six areas mentioned. A student receiving a "Satisfactory" grade in any single subject group are automatically entitled to efficient school work. In those matters, and in the prevention of emergencies during the school week which shall so equal the student's energy and interest as to make him worthy for school work, we cordially invite and confidently expect the active and explained co-operation of parents.

Superintendent *Techer*

Principal **Parents Please Sign and Return to Class Teacher**

Form 16-75B

Senior High

STUDENT _____

SUBJECT _____

TEACHER _____ Fall, Spring, 19 _____

	Period 1	Period 2	Period 3	Sem. Grade	Sem. Credits
SUBJECT MARK					
Times Absent					
Citizenship					

A check (✓) indicates need for improvement.

TEACHER'S COMMENTS:

1st _____

2nd _____

Semester _____

MEANING OF GRADES:

A, superior; B, above average; C, average; D, below average;
F, not passing; Inc., incomplete
(P, E, grades are E, S, F, only)

PARENT'S SIGNATURES:

1st Report _____

2nd Report _____

Home Room Teacher

Senior High

APPENDIX D

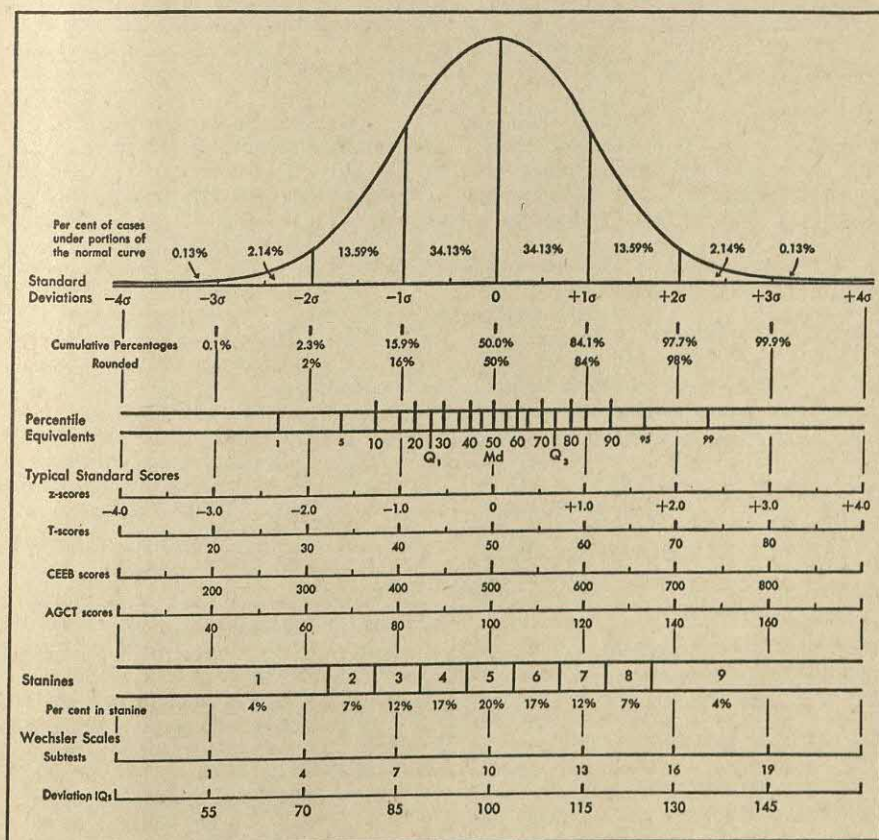
TABLE 32

Percentage of Area Under the Normal Curve Between Mean Ordinate
and Ordinate at Given z Score

z Score to Second Decimal Place										
z Score	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1514	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0.7	2580	2611	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3290	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4383	4394	4406	4418	4429	4441
1.6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4632
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2.0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2.9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3.0	4987									
3.5	4998									
4.0	49997									
5.0	4999997									

TABLE 33

Several Types of Score as Related to a Normal Probability distribution *



* The Psychological Corporation, *Test Service Bulletin*, No. 47 (September, 1954), p. 6.

INDEX

- Ability grouping, 431-432
- Achievement tests, 463-475, 480-482, 486-492
 - (*see also* specific school subjects)
- Accuracy, as a dimension of performance, 330
- Activity subjects (*see* Performance-activity areas)
- Administration of tests, 118-121
- Age norms, 159
- Algebra (*see* Mathematics)
 - example items, 321
- Anecdotal forms and procedures, 53-54
- Arithmetic (*see* Mathematics)
 - example items, 320
- Arithmetic mean, 144-149
 - comparison with median, 149-150
 - outline for computing, 148
- Arrangement of elements items
 - in general, 85, 86, 88
 - use and construction, 112
- Art
 - dimensions, 333
 - evaluation in, 356
 - example of rating chart, 345
- Attitudes and interests
 - dimensions, 390-391, 408-409
 - evaluative standards, 390-391, 414-415
 - example of an opinion scale, 285
 - forms of measurement symbols, 390-391
 - measuring procedures, 390-391, 409-414
 - self-reporting devices, 409-411
 - tests of opinion, 411-414
- Attributes (*see* Dimensions)
- Average (*see* Arithmetic mean)
- Average deviation (*see* Mean deviation)
- Ayres Handwriting Scale, 252, 472
- Behavior rating scales, 56-57, 416-418
- Beliefs, 388, 389, 391
- Bias
 - cultural and verbal bias in intelligence tests, 374-376
 - defined, 45
- Bias (*Cont.*):
 - as a factor in observational evaluations, 61
- Business education
 - dimensions, 335
 - evaluation in, 356-357
 - examples of a plan for evaluating a typing performance, 349-350
- Central tendency, measures of, 139-150
- Character and citizenship (*see* Personality)
 - defined, 387
 - dimensions, 393, 415-416
 - evaluative practices, 419
 - evaluative standards, 418-419
 - forms of measurement symbols, 416
 - measuring procedures, 416-418
- Characteristics (*see* Dimensions)
- Check lists, 52-53
 - for performance tests, 342-343, 346
- Citizenship (*see* Character and citizenship)
 - as a factor in subject grading, 275
- Classification symbols, 9-11
- Coherency, as a dimension of performance, 331-332
- Completion or fill-in items, 110, 111
- Components, as dimensions of achievement, 29
- Composition, 238-250
 - composition tests, 240-243
 - dimensions, 238-239
 - evaluative standards, 248-250
 - forms of measurement symbols, 239-240
 - marking compositions, 240
 - mechanics of grammar tests, 243-244
 - procedures for measuring, 240-248
 - spelling tests, 244-246
 - tests of rhetoric or effectiveness, 246-248
 - vocabulary tests, 246
- Concomitant variation, 173
- Constructs, 26-28
- Correlation, 172-181
 - coefficient of, 174-181
 - as concomitant variation, 173

- Correlation (*Cont.*):
 - interpretation of, 180-181
 - as a ratio, 178
- Cost of testing, 428
- Criterion
 - defined, 45
 - groups, 405
 - in validating tests, 124, 186
- Cultural bias in intelligence tests, 375-376
- Cumulative percentage curve, 156-157
- Cumulative records, 429-430
- Description as a form of measurement, 11-12
- Deviation, 144-145
- Diagnosis of disabilities
 - achievement in general, 425-426
 - arithmetic, 314-315
 - handwriting, 252
 - reading, 234-235
 - social studies, 277-278
 - speech, 352-354
 - spelling, 246
- Difficulty of test items, 100-101, 115
- Dimensions
 - art, 333
 - attitudes and interests, 408-409
 - business education, 335
 - citizenship, 393, 415-416
 - common to educational phenomena, 29-32
 - composition, 238-239
 - criteria for measurability, 20-24
 - defined, 20
 - handwriting, 251
 - home economics, 336-337
 - industrial arts, 335
 - inferred, 26-28
 - intelligence, 362-366
 - language arts, 220
 - list of educational dimensions, 17-18
 - literature, 253-255
 - mathematics, 309-313
 - music, 334-335
 - performance-activity subjects in general, 330-332
 - personality and character, 388-394
 - physical education, 337-338
 - principles in selection of, 28-29
 - reading, 228-230
 - reading readiness, 222-223
 - science, 296-299
 - social studies, 267-269
 - speech, 236
- Directions, in testing, 116, 119, 120
- Director of a testing program, 434
- Discrimination, as a dimension of performance, 331
- Discrimination in test items, 98-100
- Distributions, types of, 138
- Drawings, as personality projections, 396, 399
- Driving performance check list, 346
- Economy of effort as a dimension of performance, 331
- Educational phenomena
 - common dimensions of, 29-32
 - for which measurement may be desired, 17-18
 - preparing for measurement, 17-33
- Efficiency of a measuring procedure, 43, 44
- English, 238-250, 253-262
 - analytic evaluation of achievement, 261-262
- Error, as a dimension of achievement, 30, 330-331
- "Essay" examinations, questions, 63
(*see* Free response procedures)
- Evaluation (*see* Evaluative standards)
 - defined, 2, 190
 - as a direct outcome of observations, 50, 51
 - as distinguished from measurement, 190-192
 - evaluative standards, 192-195
 - examples of, 198-203
 - steps in, 196-198
 - symbols of, 195-196
 - (*see* specific school subjects)
- Evaluative standards (*see* Evaluation)
 - art, 356
 - attitudes and interests, 390-391, 414-415
 - business subjects, 356-357
 - citizenship, 393, 418-419
 - composition, 248-250
 - course objectives as standards, 272-273
 - criteria for, 193-194
 - distribution standards, 194
 - emotional standards, 195
 - example for history, 279-280
 - example for reading, 199, 202
 - handwriting, 252-253
 - industrial arts, 357
 - intelligence, 378
 - language arts, 221-222
 - levels of understanding as, 204
 - literature, 259-260
 - mathematics, 313-319
 - music, 356
 - nature and source of, 192-193
 - percentage standards, 194-195
 - performance-activity subjects, 354-355
 - performance scales as standards, 203-205
 - personality and character, 403-407
 - physical education, 357-358
 - reading, 235
 - reading readiness, 225, 227

- Evaluative standards (*Cont.*):
 - science, 299-301
 - used in selecting dimensions for measurement, 29
 - social studies, 272-274
 - speech, 237
 - teacher opinion as a standard, 274
 - textbooks as standards, 273
- Experience as a factor in reading readiness, 224-225
- Factorial scoring, 76-81
 - factor counting, 77-79
 - factor rating, 76-77
 - factor weighting, 79-81
- Factors (*see* Dimensions)
- Fears, 388, 389, 391
- Forms of measurement symbols
 - attitudes and interests, 390-391
 - citizenship, 393, 416
 - classification-description symbols, 9-12
 - composition, 239-240
 - criteria for selection, 13-14
 - free response tests, 73
 - guided response tests, 89
 - handwriting, 251
 - intelligence, 366-368
 - language arts, 220
 - literature, 255
 - mathematics, 319
 - observation, 50
 - performance-activity subjects, 338
 - personality and character, 394
 - product analysis, 73
 - rank symbols, 8-9
 - reading, 233-234
 - reading readiness, 223
 - scale symbols, 6-8
 - science, 301
 - social studies, 270
 - speech, 237
- Free-response procedures
 - in general, 69-83
 - applicability, 82
 - composition, 240-242
 - example in eighth grade U. S. History class, 286, 289-290
 - forms of measurement symbols, 73
 - handwriting, 251-252
 - literature, 257-258
 - mechanics, 244
 - methods of eliciting free responses, 69-73
 - methods of scoring free responses, 66-67, 73-82
 - personality and character, 396-397, 398-399
 - reading, 230
 - rhetoric, 248
 - social studies, 270-271
- Free-response procedures (*Cont.*):
 - speech, 352-354
- Frequency, as a dimension of achievement, 30
- Frequency distribution, 133-134
 - types of, 138
- Frequency intervals, 131-133
- Frequency polygon, 136-137
 - smoothed curve, 137-138
- Frequency table, 131-134
 - outline for construction, 133
- Geometry (*see* Mathematics)
 - example items, 322-324
- Grade norms, 159
- Grammar, 243-244
- Graphic rating scales, 56, 417
- Guess Who tests, 400-401
- Guessing, corrections for, 117-118
- Guidance, uses of testing in, 429-431
- Guided response procedures
 - in general, 85-89
 - administration of tests, 118-121
 - algebra, 321-322
 - arithmetic, 320
 - attitudes and interests, 409-414
 - composition, 242-248
 - construction of guided response tests, 90-118
 - defining phenomena and dimensions to be measured, 91-94
 - forms of measurement symbols, 89-90
 - geometry, 322-324
 - guessing, 116-118
 - item sequence, 114
 - literature, 256-259
 - mechanics of grammar, 243-244
 - personality and character, 397-398
 - preparation of items, 94-112
 - reading, 230-234
 - reading readiness, 225, 226-227
 - rhetoric, 246-248
 - science information, 301
 - scientific thinking, 302-306
 - scoring guided response tests, 98, 116
 - social studies, 270-272
 - spelling, 244-246
 - standardized tests, 121-126
 - test directions, 116, 119
 - test length, 113-114
 - timed tests, 115-116
 - vocabulary, 246
- Handwriting, 250-253
 - evaluative standards, 252-253
 - handwriting scales, 251-252
- Hearing, as a factor in reading readiness, 224

- Histogram, 135-136
 - outline for constructing, 135
- Home economics
 - dimensions, 336
 - example of a rating form, 350-351
- Industrial arts
 - dimensions, 335
 - evaluation in, 357
 - example of a rating form, 349
- Instrument, defined, 44
- Intellectual abilities and skills, classification of, 31
- Intelligence, 361-385
 - administering and scoring intelligence tests, 377-378
 - culture bias in measuring, 375-376
 - defined, 361-362
 - dimensions, 26, 362-366
 - dimensions commonly measured, 364-366
 - dimensions inherent in tests, 363-364
 - evaluations of, 378-379
 - group tests, 371
 - heredity and environment in, 379
 - indexes of intelligence, 366-368
 - individual tests, 371
 - list of tests, 475-479
 - measures of, related to school practice, 379-381
 - mental age (MA), 366
 - percentile ranks, 367-368
 - profiles, 371-372, 373
 - relation to reading readiness, 223-224, 227
 - reliability and validity of intelligence tests, 372-376
 - relation to school marks, 380
 - specific uses of measures of, 381
 - test items, 363-364, 369-371
 - tests, 368-378
 - theories of, 362-363
 - verbal bias in measuring, 374-375
- Intelligence quotient(s) (IQ), 367
 - deviation IQ's, 367
 - distribution in an unselected population, 168
 - parental interpretation of, 380-381
 - ratio IQ's, 367
 - unreliability in, 368
 - variability among tests, 376
- Intensity, as a dimension of performance, 331
- Interests (*see* Attitudes and interests)
- Interpercentile range, 151-152
 - comparison with standard deviation, 154
- Interpreting data, 305
- Interquartile range (*see* Quartile deviation), 152
- Intervals (*see* Frequency intervals)
- Item(s) (*see* Test items)
- Item analysis procedures, 98-101
 - difficulty, 100-101
 - discrimination, 98-100
- Job analysis, 341
- Knowledge, classification of, 31
- Knowledge and understanding
 - generalized evaluative standard, 204
 - literature, 256-257
 - mathematics, 310-312
 - science, 297-299
 - social studies, 267-269, 270-271
- Labeling items, 87, 111-112
- Language arts, 219-265
 - composition, 238-250
 - dimensions, 220
 - English in secondary grades, 260-262
 - evaluative standards, 221
 - forms of measurement symbols, 220
 - handwriting, 250-253
 - literature, 253-260
 - marking practices, 222
 - mechanics, 243, 244
 - procedures of measurement, 220-221
 - reading, 222-236
 - rhetoric, 247, 248
 - speech, 236-238
 - spelling, 244-246
 - vocabulary, 246
- Literature, 253-260
 - dimensions, 253-255
 - evaluative standards, 259-260
 - forms of measurement symbols, 255
 - literary appreciation, 257-259
 - procedures for measuring literary knowledge, 256-257
- Logical structure, 311-312, 317, 318
- Man-to-man rating scales, 417-418
- Map and chart work, 267, 270
- Marking and reporting
 - in general, 205-212
 - art, 356
 - business subjects, 356-357
 - citizenship, 275, 404, 407
 - composition, 250
 - criteria for, 210-212
 - function of school marks, 205-206
 - handwriting, 253
 - industrial arts, 357
 - language arts, 222
 - literature, 259-260
 - mathematics, 313
 - music, 356
 - parent-teacher conferences, 209

- Marking and reporting (*Cont.*):
 physical education, 357-358
 practices in, 207-210
 "psychological" marking, 275-276
 reading, 235-236
 report card examples, 494-497
 single letter marking, 207-208
 social studies, 274-276
 speech, 238
- Matching items, 86, 109
- Mathematics, 309-326
 compared with science, 294
 dimensions, 310-313
 evaluative standards, 313-319
 forms of measurement symbols, 319
 measuring procedures, 319-324
 nature of, 309-310
 standardized tests, 324-325, 480-482
- Mean, 144-149
 comparison with median, 149-150
 outline for computing, 148
- Mean deviation, 152-153
- Measurability, conditions of, 20-24
 (*see* Dimensions)
- Measurement and evaluation
 definition, 2
 factors affecting validity and efficiency, 3
- Measurement symbols (*see* Forms of),
 5-16
- Measuring procedures
 in general, 34-47
 attitudes and interests, 390-391, 409-414
 basic properties of, 35-40
 citizenship, 393, 416-419
 composition, 239-248
 criteria for, 40-44
 efficiency of, 43-44
 free-response tests, 63-83
 function of, 35
 guided response tests, 84-126
 handwriting, 251-252
 intelligence, 368-377
 language arts, 220
 literature, 256-259
 mathematics, 319-325
 observation, 48-61
 performance-activity subjects in general,
 338-344
 personality and character, 390-403
 product analysis, 63-83
 reading, 230-235
 reading readiness, 223-225
 reliability of, 42-43
 science, 301-308
 social studies, 269-272
 speech, 237
 validity of, 40-42
- Median, 140-144
 comparison with mean, 149-150
- Median (*Cont.*):
 outline for computing, 143
- Mental age (MA), 366
- Mental maturity (*see* Intelligence)
- Mode, 139
- Model (in test construction), 91
- Multiple-choice items, 86, 108, 109
- Murray pictures, sample, 71
- Music
 dimensions, 334-335
 evaluation in, 356
 example of performance test, 347-348
- Norm(s)
 age norms, 159
 defined, 45
 determining, 158-160
 gauging class progress by, 433
 grade norms, 159
 percentile norms, 159
 precautions in using, 160, 432-433
 standard score norms, 159-160
- Normal probability curve, 162-169
 applications, 166-172
 area z-score relationships, 498
 defined, 165
 related to several types of score, 499
- Objective tests, 84
 (*see* Guided response procedures)
- Objectives (*see* Dimensions)
- Observation, 48-62
 as a basic procedure of measurement,
 48-50
 anecdotal forms and procedures, 53-54
 check lists, 52, 53
 citizenship, 416-419
 devices used in, 52-59
 driver training, 346
 evaluation as a direct outcome, 50, 51
 forms of measurement symbols appropriate to, 50
 handwriting, 252
 home economics, 343
 industrial arts, 348, 349
 literature, 258
 music, 347-348
 performance-activity subjects, 338-345
 personality and character, 394-395
 physical education, 343, 351-352
 principles of validity and reliability in,
 59-61
 rating scales, 54-59
 reading, 235
 reading readiness, 224
 science, 306-307
 social studies, 270-271
 speech, 237
- Ogive, 156-157

- Open-end statements, questions, 70
- Oral language, a dimension of reading readiness, 224
- Paired comparison items, 412
- Peer ratings, 399–403
- Percentile, 151–152, 156–157
- Percentile rank, 156–157
 - comparison with standard score, 158
 - norms, 159
- Performance-activity areas, 329–360
 - dimensions, general, 330–332
 - evaluative standards, 354–358
 - examples of measuring procedures, 344–354
 - forms of measurement symbols, 338
 - measuring procedures, 338–354
 - performance test, 339–344
 - process and product, 329–330
- Performance tests, 339–354
- Personality and character, 386–423
 - in general, 386–407
 - analysis of drawings, 396
 - attitudes, 390, 407–415
 - character attributes, 393
 - clinical definitions of maladjustment, 406–407
 - criterion groups, 405
 - defined, 387
 - dimensions, 388–394
 - evaluative practices, 407
 - evaluative standards, 403–407, 414–415
 - fears, beliefs, and values, 388–389, 391
 - forms of measurement symbols, 390–393, 394
 - Guess Who tests, 400–401
 - interests, 390–391, 407–415
 - measuring procedures, 390–393, 394–403
 - opinion tests, 397–398
 - outline of dimensions, measurement forms, procedures, and evaluative standards, 390–393
 - peer ratings, 399–403
 - personality structure, 389, 392, 394
 - projective techniques, 71, 398–399
 - sociograms, 401–403
 - standardized tests, 483–486
- Personality structure, 389, 392, 394
 - personality traits, 389, 392, 394
 - personality types, 389, 392
 - polar dimensions, 389, 392
- Physical education
 - dimensions, 337
 - evaluation in, 357–358
 - example of a plan for evaluating diving, 351
- Preparing phenomena for measurement, example of, 91–94
- Probability, 163–164
- Problem questions, 72
- Problem solving
 - as a factor in intelligence, 364, 365, 369, 370
 - mathematics, 312–313, 317–319
 - science, 297–299, 302–306
 - social studies, 269, 271
- Product analysis
 - in general, 69–83
 - applicability, 82
 - art, 345–346
 - clothing, 350–351
 - composition, 240
 - forms of measurement symbols pertinent, 73
 - handwriting, 251–252
 - mechanics, 244
 - methods of scoring products, 64–65, 73–82
 - personality and character, 395–396
 - rhetoric, 248
 - science, 306–307
 - social studies, 270
 - spelling, 244
 - typing, 348, 349
 - vocabulary, 246
- Product scales, 75–76
- Profile(s), 125–126, 171–172, 371–372, 373
- Projective techniques, 71, 398–399
- Properties (*see* Dimensions)
- Provide-an-answer items
 - in general, 85, 87
 - use and construction, 110–112
- “Psychological” grading or marking, 275–276
- Publishers’ addresses, 493
- Qualities (*see* Dimensions)
- Quantitative thinking, 311
- Quartile deviation (Q), 152
- Questionnaires (*see* Self-reporting procedures)
- Range, 150–151
- Rank order, 155
- Rank symbols, 8–9
- Ranking products and free responses, 74–75
- Rate, as a dimension of achievement, 30
- Rating products and free responses, 74
- Rating scales
 - in general, 54–59
 - category scales, 55
 - in citizenship, 416–418
 - continuum scales, 55, 56
 - criteria for design and use, 59
 - graphic or descriptive scales, 56, 417
 - illustrated, 56–57
 - man to man, 417

- Rating scales (*Cont.*):
 measurement symbols derived from, 58
 number of categories or intervals in, 58
 for performance tests, 343, 345, 349,
 350–351, 352, 353
 with unequal intervals, 57–58
- Reading, 228–236
 diagnosing disabilities, 234–235
 dimensions, 228–230
 evaluative standards, 235–236
 measuring procedures in general, 230
 reading test scores, 233–234
 reading tests, 230–234
- Reading readiness, 222–227
 dimensions of, 222–223
 evaluative standards, 227
 forms of measurement symbols, 223
 measuring procedures, 223–225
- Reading readiness tests, 225–227
 correlation of scores with teacher ratings,
 225
 illustrated, 226–227
- Recognition tests, 339–340
- Recordings
 used as standards in music, 356
 used as standards in speech, 237
- Referral for individual setting, 430–431
- Regression line, 177–180
- Relative position, 155–158
- Reliability, 42–43, 181–186
 coefficient of, 43, 185
 definition of, 42, 181–182
 fisherman's ruler analogy, 42–43
 interpretation of, 185–186
 sources of inaccurate measurements, 182
 standard error of measurement, 185
 of standardized tests, 124
 statistical procedures for determining,
 183–185
- Report cards, 205–212
 examples, 494–497
- Representative measures, 139–150
- Rhetoric, 246–248
- Rorschach ink blot, sample, 71
- Sampling
 categorical, 39
 related to guided response procedures,
 102–104
 related to measuring procedures, 38–40
 temporal, 40
- Sampling error, 169–170
- Scale symbols, 6–8
- Scatter diagram, 174–175
- School marks (*see* Marking and reporting)
- School-wide testing programs, 424–438
 ability grouping, 431–432
 administering tests, 427
 cost, 428
- School-wide (*Cont.*):
 cumulative records, 429, 430
 director of a testing program, 434
 focal points of, 424–425
 frequency of testing, 432, 434–435
 guidance, 429–431
 handling results, 427–428
 instructional facilitation, 432–433
 instruments and procedures, 425–429
 locally devised tests, 428
 ordering tests, 427
 reasons for, 424
 referral for individual testing, 430–431
 selection of tests, 426
 scope of testing, 434–435
 scoring tests, 427
 tenets of an efficient testing program,
 433–435
 uses of, 429–433
 vocational guidance, 430
- Science, 295–308
 dimensions, 296–299
 evaluative standards, 299–301
 forms of measurement symbols, 301
 instruction in schools, 295–296
 measuring procedures, 301–308
 nature of science, 295–296
 relation with mathematics, 294
 special measurement problems, 308
 standardized tests, 307–308, 489–490
- Scientific thinking, 302–305
- Scoring
 defined, 45
 electrical or machine, 98, 116
 factor counting, 77–79
 factor rating, 76–77
 factorial scoring, 76–81
 free response ("essay") tests, 66, 67, 73–
 82
 guided response tests, 98, 116
 keys, 98, 116
 products, 64–65, 73–82
 ranking products and free responses,
 74
 rating products and free responses, 74
 self-reports, 410–411
 in testing programs, 427
 tests of opinion, 414
 use of a product scale, 75–76
 weighting in, 79–81
- Selection of an answer item
 in general, 85, 86
 use and construction, 106–110
- Self-reporting procedures, 396–397, 409–
 411
- Sensory data, 24
- Short answer items, 87, 111
- Simulated task or performance
 as a basic measurement concept, 41

- Simulated task (*Cont.*):
 as a test procedure in performance-activity subjects, 340
- Social distance, 411, 412
- Social studies, 266-292
 evaluative standards, 272-274
 forms of measurement symbols, 270
 marking and reporting, 274-276
 measuring procedures in general, 270-272
 phenomena and dimensions in general, 267-270
 social studies units, 276-277
- Sociograms, 401-403
- Sociometry, 399-403
- Speech, 236-238
 dimensions, 236
 evaluative standards, 237-238
 forms and procedures of measurement, 237
 marking achievement in, 238
 recordings as evaluative standards, 237-238
- Speed, as a dimension of performance, 30, 330
- Spelling, 244-246
- Standard analysis systems, 37-38
- Standard deviation, 153-154
 comparison with interpercentile range, 154
 outline for computing, 154
- Standard differential responses, 36-37
- Standard error of measurement, 169-172, 185
 as related to reliability coefficient, 185
 as related to score profiles, 172
- Standard score, 157-158
 comparison with percentile rank, 158
 norms, 159
- Standard stimulations, 35-36
- Standardized, defined, 44-45
- Standardized tests and testing
 in general, 121-126
 ability-grouping, 431-432
 achievement batteries, 463-467
 annotated bibliography, 463-492
 art, 467-468
 attitudes and interests, 414
 business, 468-469
 cost of, 428
 English, 470-471
 evaluating standardized tests, 123-125
 foreign language, 472
 guidance, 429-431
 handling results, 427-428
 handwriting, 251-252, 472
 health, 473
 home economics, 473-474
 instructional facilitation, 432-433
- Standardized tests (*Cont.*):
 intelligence, 368-377, 475-479
 language arts, 221
 literature, 256-257
 mathematics, 324-325, 480-482
 music, 482
 ordering, 427
 personality and character, 397-399, 483-486
 reading, 231-234, 486-489
 reading readiness, 225, 226-227
 restrictions on use, 433, 435
 science, 307-308, 489-491
 scoring, 427
 selection of, 426
 social studies, 271, 272, 491-492
 sources, 122-123
 spelling, 245-246
 uses of, 429-433
 vocabulary, 246
 wood working, 475
- Stencil key, 98, 117
- Stimulus pictures, 71-72
- Stimulus words and objects, 70-71
- Stories to complete, 72
- Strip key, 117
- Study habits, 286-289
 observation schedule for, 287-288
- Tabular and graphical portrayal of measurements, 130-139
- Taxonomy of Educational Objectives*, 31
- Test(s)
 in various subjects (*see* subject entries)
 attitudes and interests, 409-414
 defined, 44
 free-response tests, 69-83
 Guess Who tests, 400-401
 guided response tests, 85-126
 intelligence tests, 368-377
 performance tests, 339-354
 personality-character, 396-399
 projective tests, 71, 398-399
 recognition tests, 339-340
 standardized tests, 121-126, 463-492
 testing programs, 424-436
 tests of opinion, 411-414
 work sample tests, 340-344
- Test items
 chance factor of, 104, 105
 defined, 44
 difficulty, 100-101, 104, 105, 115
 discrimination, 98-100
 format for responses, 98
 function of, in guided response tests, 94-96
 language of, 96-97
 need for independence, 101-102
 sampling function, 102-104

- Test items (*Cont.*):
 - sequence of, 114
 - types of guided response items, 104-112
- Time, as a dimension of achievement, 29-30
- Timed tests, 115-116
- Timing, as a dimension of performance, 331
- Trend line, 177-180
- True-false items, 86, 107-108
- Understanding (*see* Knowledge and understanding)
- Understanding, levels of
 - example in mathematics, 314-318
 - example in science, 300
 - example in social studies, 279-280
 - as a generalized evaluative standard, 204
- Validity
 - defined, 40-41
 - factors bearing on, 41-42
 - of standardized tests, 124-125
 - statistical estimates of, 186-187
- Values, 388-389, 391
- Variability of groups, measures of, 150-155
 - interpercentile ranges, 151-152
 - mean deviation, 152-153
 - quartile deviation, 152
 - range, 150-151
 - standard deviation, 153-154
- Variance
 - defined, 153
 - error, 178
 - explained, 178
 - as related to correlation, 177-181
 - total, 177
- Vision, as a factor in reading readiness, 224
- Vocabulary, 246
- Vocational guidance, 430
- Weighting
 - difficulty weighting, 80-81
 - importance weighting, 79-80
 - in scoring products and free responses, 79-81
- Work sample test, 340-344
- z-score (*see* Standard score)

